a cura di:
Manuel Barbera
Elisa Corino
Cristina Onesti

corpora e linguistica in rete

a cura di:
Manuel Barbera
Elisa Corino
Cristina Onesti

corpora e linguistica in rete

Divide et adnota!
Julii ficti Caesaris *De bello grammatico*.

0. Indice.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. iij-iv.

PREMESSA

CARLA MARELLO
(Università di Torino)

j. L'italiano ed altre lingue nella varietà dei corpora. *Una introduzione.*

0. PREMESSA. *L'italiano nella varietà dei testi* è la parte iniziale del titolo della ricerca[1] da cui è scaturito il cuore degli studi raccolti in questo libro. Variare con la menzione delle altre lingue è doveroso nel titolo di questo scritto introduttivo perché il gruppo di ricerca ha approntato corpora non solo per l'italiano, ma anche per francese, inglese, spagnolo e tedesco; ed altre lingue ancora sono in lavorazione (cfr. Barbera ¶ 1, p. 7 n. 10). La varietà dei testi ha determinato la grande varietà testuale dei corpora preparati soprattutto per l'italiano: si va dagli scritti accademici dell'Athenaeum Corpus, ai molteplici registri linguistici presenti nei newsgroup di NUNC (Newsgroup UseNet Corpora), all'italiano di apprendenti stranieri in VALICO (Varietà di Apprendimento Lingua Italiana Corpus Online) e di studenti italiani in VINCA (Varietà di Italiano di Nativi Corpus Appaiato), all'italiano duecentesco del Corpus Taurinense, che è servito come durissima palestra di allenamento per tutti gli altri.

Il libro riproduce parte del programma del convegno internazionale "Corpora e linguistica in rete" tenutosi a Torino il 30 settembre 2005, ma non si può dire che ne costituisca gli atti. Da una parte perché vorrebbe disegnare un progetto organico, raccogliendo anche contributi a quel convegno precedenti (ad es. Allora - Barbera ¶ 5 e Barbera ¶ 8) e successivi (ad es. Barbera - Corino - Onesti ¶ 3 e Barbera ¶ 23). Da un'altra parte perché molte delle ricerche che allora vennero presentate e discusse in vista del termine triennale del menzionato FIRB hanno frattanto potuto beneficiare della proroga del progetto fino alla primavera del 2007 e sono state quindi ulteriormente approfondite e sviluppate; alcune linee di ricerca, anzi, hanno tratto spunto proprio dalle discussioni del convegno.

Rimando all'indice estremamente dettagliato che si trova al ¶ 26, in fondo al volume, per una panoramica dei saggi che il libro contiene e dedico invece queste pagine introduttive a mettere in rilievo i principali punti di forza della ricerca che è organicamente[2] qui presentata per la prima volta.

1. *META-CORPUS LINGUISTICS.* Le molteplici varietà, di lingua e di testi, hanno trovato nella formazione in filologia, in linguistica testuale ed in linguistica computazionale dei ricercatori del gruppo terreno fertile per innescare una serie di riflessioni approfondite su che cosa significhi fare corpora elettronici, metterli a disposizione ed interrogarli. Eminentemente metalinguistici sono tutti i contributi raccolti nella prima parte del volume, articolata tematicamente, ma questo interesse non è assente neppure nella seconda parte, articolata in base ai corpora od alle basi dati testuali prese come punto di partenza.

1.1 ASPETTI LEGALI. Il punto di partenza, attualissimo ma nient'affatto scontato, per queste riflessioni è stato quello legale: problema, questo, molto avvertito nella comunità della linguistica dei corpora, ma in genere ritenuto disturbo vitando. Qui, invece, non solo il punto non è

[1] *L'italiano nella varietà dei testi. L'incidenza della variazione diacronica, testuale e diafasica nell'annotazione e interrogazione di corpora generali e settoriali*: progetto FIRB RBAU014XCF 2001, coordinatore Carla Marello.
[2] Sono in preparazione altre pubblicazioni che approfondiranno settori specifici della ricerca; e molti articoli di ricercatori del gruppo sono già apparsi in riviste, atti di convegno ed opere collettive, pubblicati in Italia e all'estero: per una panoramica complessiva si veda oltre Barbera ¶ 1.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. vij-xij.

stato evitato, ma crediamo anzi di averne proposto una possibile soluzione. Proficua e necessaria è stata naturalmente la collaborazione con esperti legali interessati ai problemi del diritto d'autore relativamente a banche dati ed altre opere collettive (ed alla loro pubblicazione in rete), che ha dato vita ad una sezione non piccola della prima parte del libro.

Quest'aspetto legale, va tra l'altro rimarcato, è ancor più vitale per chi, usando denaro pubblico ed operando all'interno di università e centri di ricerca pubblici, voglia rendere non solo di comune dominio i propri prodotti, ma li voglia anche mettere gratuitamente a disposizione della comunità.

Il volume contiene, così, i modelli dei primi contratti di tipo *Creative Commons Public Licences* per i corpora.

1.2 ASPETTI TECNICO-DEFINITORI. La prima parte degli interventi della prima sezione è dedicata a quello che propriamente chiameremmo *metalinguistica generale dei corpora*, e cioè alla definizione di che cosa sia un corpus elettronico, di quali siano le sue caratteristiche individuanti, e di come poi lo si assembli ed infine interroghi con appropriati e appositi programmi.

La definizione puramente architettonica, *eine Art Scheingesims* (per usare l'immagine wittgensteiniana posta in epigrafe a Barbera - Corino - Onesti ¶ 3), di *corpus* è stata, curiosamente, perlopiù finora elusa nella letteratura tecnica, ma era indispensabile per poter impostare un discorso legale che non fosse edificato sulla sabbia. Un vantaggio ulteriore di ciò è stato quello di meglio svincolare il discorso storiografico sulla *corpus linguistics* dalla specifica natura dei corpora, rendendolo più neutrale, e facilitando così lo sgancio dalla tradizione esclusivamente anglistica (che vede in Fries il grande *generis auctor*) e l'aggancio alla tradizione italiana, dalla prima Crusca fino al padre Busa, così efficacemente proposto da Francesco Sabatini (cfr. soprattutto Sabatini 2006), che anzi ne radica i semi nelle scaturigini stesse della storia della lingua italiana (cfr. qui Sabatini ¶ ij). Un vantaggio ulteriore è quello di liberalizzare l'uso, da parte del linguista di corpora, anche di altri strumenti oltre a quello dei corpora veri e propri, senza detrimento o compromissione della propria disciplina: alcune delle applicazioni nella parte seconda del volume dànno chiaro esempio di ciò, spaziando attraverso diverse basi di dati testuali come nei contributi di Korzen e Conte.

Importante, in questo àmbito, è anche il contributo chiarificatore alla definizione dei concetti di *token* e *type*, riportati alle loro fondazioni semiotiche e filosofiche (Peirce, Quine), normalmente omesse, od ignorate, non senza conseguenze teoriche e talora anche pratiche.

Questo nucleo tematico si conclude appropriatamente con un importante contributo sulla costruzione e rappresentazione informatica dei corpora in CWB (Corpus Work Bench), nonché sulla loro interrogabilità, tanto dal punto di vista della sintassi di interrogazione, quanto da quello delle interfacce web per gli utenti.

1.3 ASPETTI TESTUALI. Un'altra novità del volume è la consistente presenza di studi testuali, laddove di solito la linguistica testuale è invece scarsamente rappresentata in *corpus linguistics*: impostato come si è fatto l'assetto legale dei corpora, e consentendo l'accesso ai testi interi presenti nei corpora, si è reso così davvero possibile fare linguistica testuale con i corpora. A ciò hanno congiurato vuoi gli interessi testuali di molti degli studiosi del gruppo, vuoi lo stimolo offerto dai particolari materiali che costituiscono i NUNC (cui è monograficamente dedicato Corino ¶ 13, e per cui cfr. anche oltre, § 2.3).

Si è infatti quasi subito innestata prepotentemente nella discussione la questione del modo di produzione dei testi che costituiscono i NUNC: si tratta infatti di testi prodotti in rete, pensati per un peculiare tipo di lettura e fruizione. E la rete come mezzo per accedere a tutti i tipi di testi elaborati e immessi nei corpora presenti in corpora.unito.it ha creato un ulteriore filone di indagine rivelatosi di primaria importanza.

2. SVILUPPI DELLA RICERCA. La corposa seconda sezione del libro, costituita da studi, anche di rilievo metalinguistico, ma che partono sempre da specifici corpora od altri materiali, ed organizzata, come s'è detto, in base ad essi (percorrendo un sentiero che partendo dal Corpus Taurinense, attraversa l'Athenaeum, sosta sui NUNC, e poi, dopo avere lambito Vinca, approda ai "non-corpora", dalla base dati testuali di Mr. Bean ai testi della Costituzione svizzera), è conformemente percorsa da almeno due anime: da un lato sì discutere e risolvere problemi connessi a vari tipi di annotazione (POS-tagging, disambiguazione, ecc.), ma dall'altro anche mostrare in studi significativi le caratteristiche dei corpora approntati dal gruppo di ricerca e consultabili in corpora.unito.it.

2.1 *CASE STUDIES*. Agli autori degli studi su specifici fenomeni linguistici, sia agli interni al gruppo di ricerca sia agli esterni, si è infatti chiesto di saggiare la facilità di interrogazione dei corpora e la significatività dei risultati che ottenevano. In particolare si è chiesto di rendere il più possibile esplicita l'entità dell'aiuto che l'avere a disposizione un corpus elettronico, preparato nel modo in cui sono preparati quelli inseriti nel nostro sito, può dare al ricercatore.

Ciascuno ha perciò cercato nei corpora i fenomeni che stava già studiando con esempi d'uso raccolti tradizionalmente od estratti da altri corpora. Fra gli scritti prodotti da ricercatori interni alcuni riflettono sulle peculiarità della lingua nella comunicazione mediata dal computer nei newsgroup ed altri casi di studio affrontano specifiche questioni come i connettivi, le collocazioni, le comparative prototipiche, gli anglismi, la negazione, gli usi deontici e anankastici di *dovere*[3] (ma su quest'ultimi torneremo tra poco: § 2.4), ecc.

2.2 LA STANDARDIZZAZIONE DEI TAGSET ED OLTRE. Volendo utilizzare lo stesso insieme di annotazioni per testi (e corpora) di lingue diverse, e tanto vari nel tempo, nel registro, nell'argomento, è stato giocoforza soffermarsi sul problema della standardizzazione, studiando, in particolare, un insieme di annotazioni morfosintattiche per parte del discorso (POS-tagset) e di articolazione interna del testo in paragrafi (markup) che potesse valere per tutte le lingue e per tutti i testi. Importante corollario è stato poi risolvere le questioni della disambiguazione degli omonimi per rendere migliori i risultati delle interrogazioni, argomento qui presente col lavoro di Tomatis, tanto più rilevante nella scarsità bibliografica in materia.

La zona più avanzata di queste ricerche è quella sui tagset ed è su questa che più il volume si sofferma, seguendo il solco della ricerca internazionale da EAGLES (Expert Advisory Group on Language Engineering Standards) ad ISLE (International Standards for Language Engineering), ed articolando il proprio discorso in due lavori, separati tra di loro da sei non inattivi anni: da un lato Barbera ¶ 8 descrive approfonditamente e diffusamente come debba essere strutturato un tagset gerarchico, appoggiandosi al tagset italiano antico costruito per il Corpus Taurinense, di cui si fornisce la descrizione di riferimento; dall'altro Barbera ¶ 23 (posto in appendice, a mo' di documentazione) muove decisamente in direzione interlinguistica, fornendo un prezioso mapping tra i molti tagset (francese, inglese, italiano moderno ed antico, spagnolo e tedesco, tutti disponibili come *parameter file*s per il Tree Tagger) attualmente usati su bmanuel.org e corpora.unito.it, tabulando tra l'altro l'ultima versione del nostro tagset per lo spagnolo (appena presentata in Barbera 2007 *i.s.*; ed il nuovo tagset per l'italiano moderno è dietro l'angolo!), e discutendo gli ulteriori principi, teorici e pratici, che ci stanno guidando nella costruzione di una suite di tagset armonizzati per le ricerche multilinguistiche.

---

[3] Siamo grati a Amedeo G. Conte che ha rielaborato, per includerla nel libro, una sua inedita ricerca precedente. Distribuito il giorno del convegno internazionale "Corpora e linguistica in rete", il suo fertile scritto sta incoraggiando ulteriori indagini nel corpus Jus Jurium.

Del grande lavoro di armonizzazione del markup testuale e dei metadata, invece, si gettano qui le sole basi teoriche (in Barbera - Corino - Onesti ¶ 3, § 1.4) in attesa di presentarne le applicazioni in altra sede.

2.3. *UMGANGSSPRACHE* AL COMPUTER. Il termine *Umgangssprache,* di spitzeriana memoria,[4] ci è parso ben adatto per indicare la varietà di lingua più largamente rappresentata nell'insieme di corpora allestiti dal gruppo, cioè quella dei newsgroup, perché ci permette di scavalcare la discussione lingua scritta vs. lingua parlata per sottolineare lo scopo della comunicazione. Se da un lato aver raccolto gruppi di discussione è anche la conseguenza della difficoltà di reperire larghe quantità di testi scritti da mettere in rete alle condizioni legali da noi volute, dall'altro si può dire che si è trattato di una circostanza felice, in quanto permette alle comunità degli studiosi di italiano, francese, inglese, spagnolo e tedesco (ed altre lingue stanno per essere raggiunte dal progetto) di analizzare un tipo di lingua scritta molto moderna, avvicinabile al parlato, però scritta dagli autori dei messaggi e, come tale, comunque manifestazione di pianificazione del testo.

La fresca contemporaneità degli esempi tratti da questo insieme di corpora non mancherà di attirare l'attenzione di chi si occupa di ricerca in morfologia e sintassi, in glottodidattica, e di chi è interessato a documentare l'acclimatazione di prestiti e di neologismi al di fuori dei giornali, nella lingua usuale di chi scrive per comunicare con altri che condividono i suoi interessi.

Non è fortuita coincidenza che il libro inauguri il ramo "Strumenti" della collana "L'officina della lingua"[5]. Si vuol sottolineare l'importanza di fare linguistica partendo da una documentazione ampia della lingua d'uso: in questo senso la metafora dell'officina richiama il luogo in cui si creano strumenti con cui si fabbricheranno prodotti che a loro volta avranno un'utilità per professionisti delle lingue: linguisti, insegnanti, giornalisti, traduttori.

2.4 DALLA TESTUALITÀ ALLA SEMANTICA. L'interrogabilità a contesti illimitati (frutto della nostra accorta politica legale) non solo ha reso accessibile la linguistica dei corpora ai testualisti, ma la ha anche portata in zone ancora più lontane e, tradizionalmente, estranee alla disciplina, come la semantica: cammino che ben si percorre dal lavoro di Angela Ferrari e del suo gruppo, a quello su "mica" di Jacqueline Visconti, a quello su "dovere" di Marco Carmello.

Anzi Carmello ¶ 21 teorizza anche questa caratteristica dei nostri corpora, e porta il suo discorso sul limine della logica deontica, preparando in ciò il terreno al contributo finale di Amedeo Conte.

3. RINGRAZIAMENTI. Desidero ringraziare *in primis* Manuel Barbera ed Ulrich Heid: il primo perché è stato fondamentale per la ricerca e per il gruppo di ricerca, come la curatela di questo volume ed i suoi contributi in esso dimostrano; il secondo perché ha generosamente collaborato fin dalla stesura del progetto e poi ha seguito passo passo gli aspetti linguistico-computazionali. Al suo Istituto, l'IMS di Stuttgart, la linguistica computazionale torinese è da tempo legata da debito di gratitudine per la concessione del CWB, il software di interrogazione di corpora, fin da tempi in cui il suo rilascio sotto GPL era ancora inimmaginabile. Ricerche di questa durata ed ampiezza non si possono portare avanti senza il sostegno delle istituzioni: oltre al Ministero dell'Università e della Ricerca, che è stato il principale finanziatore, questo progetto è stato sostenuto dall'allora Rettore dell'Università di Torino, Rinaldo Bertolino. Preziosa è stata la collaborazione del personale e dei docenti del Dipartimento di Scienze letterarie e filologiche,

[4] Memoria rinverdita dalla recente traduzione del suo *Italienische Umgangssprache* (1922) a cura di Claudia Caffi e Cesare Segre: cfr. Spitzer 1922/2007.

[5] L'altro ramo della collana, diretta da Carla Marello, "Formazione insegnanti Italiano lingua straniera", è da tempo attivo.

della Facoltà di Lingue e letterature straniere, dell'Ufficio Stampa dell'Università degli Studi di Torino e di Dario Cantino, direttore de L'Ateneo, per i testi del corpus accademico, del Centro ReTe dell'Università di Torino, che mantiene il server del sito corpora.unito.it.

Fra quanti presero la parola, o mandarono interventi, il 30 settembre 2005 desidero ringraziare il Rettore Ezio Pelizzetti, Rinaldo Bertolino, all'epoca Rappresentante generale della CRUI a Bruxelles, Mauro Massulli, dirigente Ufficio FIRB del Ministero dell'Università e della Ricerca, Ferdinando D'Isep, direttore del Centro ReTe dell'Università di Torino, Federico Reviglio del quotidiano La Stampa di Torino, altro importante "fornitore" di testi.

I colleghi Bice Mortara Garavelli; Fernando Martinez de Carnero Calzada, Livio Gaeta, Francesca Geymonat ed Elisabetta Soletti hanno disciplinato ed al contempo animato le discussioni durante il convegno e in preparazione di esso; Michele Cortelazzo (Padova), Emanuela Cresti (Firenze), Massimo Moneglia (Firenze), John Osborne (Chambéry), Davide Ricca (Torino), Salvatore C. Sgroi (Catania), Jacqueline Visconti (Birmingham e Genova) ed Ugo Volli (Torino) hanno preso parte alla tavola rotonda "Corpora elettronici come fine e come mezzo". Alcuni fanno parte del gruppo di ricerca, altri hanno seguito e seguono il nostro lavoro, facendoci profittare del confronto con il loro.

Al Presidente dell'Accademia della Crusca, Francesco Sabatini, che ci ha sempre sostenuto con attento e partecipe consiglio, tutta la nostra gratitudine per aver aperto il convegno internazionale ed aver accettato di aprire anche questo volume.

Un ringraziamento speciale a Marco Ricolfi e Marco Ciurcina che hanno prestato paziente attenzione alle nostre esigenze legali ad Amedeo G. Conte, non solo per i suoi consigli scientifici, ma anche per l'immagine che costituisce la copertina di questo libro e, rielaborata, la copertina del sito in rete.

I collaboratori stranieri del progetto, Angela Ferrari, Iørn Korzen, Ulrich Heid, Jacqueline Visconti sono stati preziosi per lo sviluppo della ricerca ben al di là dei loro contributi in questo libro. Infine le figure di Elisa Corino e Cristina Onesti, co-curatrici di questo volume, mi rammentano il graditissimo dovere di ringraziare i molti studenti di corsi quadriennali, triennali, biennali, di dottorato, che hanno capito che cosa vuol dire fare linguistica dei corpora facendola e permettendo a noi di affinare le nostre capacità di formatori in questo àmbito. Fra i risultati più cospicui e, spero, duraturi di questa ricerca c'è stato un felicissimo periodo di continua interazione fra ricerca e didattica universitaria.

BIBLIOGRAFIA.

ALLORA - BARBERA

¶ 5 Adriano Allora - Manuel Barbera, *Il problema legale dei corpora. Prime approssimazioni*, in questo volume, pp. 109-118.

BARBERA

¶ 1 Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in questo volume, pp. 3-20.

¶ 8 Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in questo volume, pp. 135-168.

¶ 23 Manuel Barbera, *Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra* guidelines *e prolegomeni*, in questo volume, pp. 373-388.

2007 *i.s.* Manuel Barbera, *I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo*, in "Cuadernos de filología italiana" XIV (2007) 11-32 in corso di stampa.

BARBERA - CORINO - ONESTI

¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

CARMELLO

¶ 21 Marco Carmello, *“Dovere” deontico e “dovere” anankastico fra semantica e pragmatica. Una ricerca* corpus-based, in questo volume, pp. 347-362.

CORINO

¶ 13 Elisa Corino, *NUNC (Newsgroup UseNet Corpora). Questioni metodologiche ed aspetti della testualità*, in questo volume, pp. 225-252.

CORINO - MARELLO - ONESTI

2006 *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 | Proceedings of the XII EURALEX International Congress. Torino, Italia, 6th-9th September 2006*, a cura di Elisa Corino, Carla Marello e Cristina Onesti, 2 voll., Alessandria, Edizioni dell’Orso, 2006.

SABATINI

2006 Francesco Sabatini, *La storia dell’italiano nella prospettiva della* corpus linguistics, in CORINO - MARELLO - ONESTI 2006, pp. 31-37.

¶ ij Francesco Sabatini, *Storia della lingua italiana e grandi corpora. Un capitolo di storia della linguistica*, in questo volume, pp. xiij-xvj.

SPITZER

1922/2007 Leo Spitzer, *Italienische Umgangssprache*, Bonn, Kurt Schroeder, 1922. Versione italiana: *Lingua italiana del dialogo*, a cura di Claudia Caffi e Cesare Segre, traduzione di Livia Tonelli, Milano, il Saggiatore, 2007.

CORPORA, STRUMENTI E SITI DI RIFERIMENTO.

Athenaeum Corpus	`http://www.bmanuel.org/projects/at-HOME.html`
bmanuel.org	`http://www.bmanuel.org`
Centro ReTe	`http://www.rete.unito.it/`
corpora.unito.it	`http://www.corpora.unito.it/`.
Corpus Taurinense	`http://www.bmanuel.org/projects/ct-HOME.html`
CWB	`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`
EAGLES	`http://www.ilc.cnr.it/EAGLES96/home.html`
IMS Stuttgart	`http://www.ims.uni-stuttgart.de`
ISLE	`http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm`
Jus Jurium	`http://www.bmanuel.org/projects/ju-HOME.html`
NUNC	`http://www.bmanuel.org/projects/ng-HOME.html`
Tree Tagger	`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`
VINCA	`http://www.bmanuel.org/projects/vn-HOME.html`

FRANCESCO SABATINI
(Università di Roma Tre)

ij. Storia della lingua italiana e grandi corpora. *Un capitolo di storia della linguistica.*

0. PREMESSA. I linguisti che oggi conducono ricerche sulla base di *corpora* testuali vivono generalmente nella convinzione che una simile impostazione del loro lavoro sia il frutto di una svolta abbastanza recente nella speculazione teorica: la svolta, legata sostanzialmente alle correnti strutturaliste, che induce a porre una base empirica – documentata ed identificabile, decisamente chiusa od anche estendibile – alla descrizione dell'intero sistema della lingua studiata. Mia intenzione, in questa sede, non è quella di discutere in generale dei vantaggi e dei limiti di questa impostazione degli studi linguistici. Dirò subito che l'utilizzazione dei corpora fornisce comunque uno straordinario arricchimento alla conoscenza dei tratti costitutivi dell'italiano, la cui descrizione è generalmente basata, nella maggior parte delle grammatiche e dei dizionari correnti fino ad anni recentissimi, su schemi assolutamente tradizionali, che nulla dicono, ad esempio, sulla struttura argomentale dei verbi, sui trasferimenti di molti elementi dalla pura frasalità alla testualità, sulle "collocazioni": tutti fenomeni che possono essere individuati con precisione e misurati nella loro estensione e distribuzione solo in base ad ampi corpora, costituiti su appropriate tipologie testuali, e non solo sincronici, ma (per le ragioni che adduco in conclusione) sufficientemente diacronici.

1. TRADIZIONE GRAMMATICOGRAFICA E LESSICOGRAFICA ITALIANA. Il mio intento principale, però, è ora un altro.

Vorrei segnalare, soprattutto a chi meno si occupa di storia linguistica italiana, che il fare preciso ricorso ad un *corpus di testi*[1] è una costante nell'intera nostra tradizione grammaticografica e lessicografica e, in termini ancora più ampi, nella storia delle dispute linguistiche fin dall'epoca di Dante. Una costante che trova la sua ragion d'essere in una condizione particolare, solitamente considerata penalizzante, della nostra lingua: la sua nascita attraverso l'opera di scrittori e la sua lunga permanenza in vita attraverso l'uso scritto, e quindi grazie al continuo sostegno dato da un canone di autori. Richiamerò qui velocemente le tappe principali di questa vicenda.

1.1 LINGUA E GRAMMATICOGRAFIA DA DANTE AL BEMBO. È noto che dietro la dimostrazione che Dante vuol dare, nel *De vulgari eloquentia*, dell'esistenza ed addirittura del primato, in campo neolatino, del *vulgare latium* (lingua colta comune all'Italia intera), c'è la sua conoscenza delle grandi sillogi dei poeti illustri ("siciliani", siculo-toscani e stilnovisti), costituite alla fine del Duecento in Toscana e chiaramente circolanti nel suo ed in altri ambienti[2]. Sono quelle sillogi (corposissime, ordinatissime) a dare il primo fondamento all'identità della lingua italiana: non strettamente situabile in un solo ambito geografico, altamente elaborata dalla pen-

[1] "Corpus di testi" corrisponde qui a quello che nella terminologia stretta di Barbera - Corino - Onesti ¶ 3, § 1.2 , ed in genere in questo volume, è "precorpus" [M.B.].

[2] Si tratta dei famosi *Canzonieri* dei quali abbiamo oramai edizioni e riproduzioni che forniscono la più ampia informazione, dalle avalliane *Concordanze della lingua poetica italiana delle origini* (cfr. Avalle 1992) ai recenti *Canzonieri della lirica italiana delle origini* (cfr. Leonardi 2000-01).

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. xiij-xvj.

na degli scrittori, quindi rintracciabile facilmente solo in un corpus di testi scritti, sia pure letti secondo i gusti di chi lo compila o lo utilizza.

Durante il corso del Trecento e del Quattrocento, nell'Italia che continua ad essere mosaico di corti e città, l'essenziale unità della lingua è sempre affidata alla circolazione di sillogi di testi esemplari, nutrite abbondantemente dalle opere dei grandissimi, ormai canonizzati, e da una crescente produzione di altri testi di notevole livello. Episodio di prima grandezza è quello della confezione (a Firenze, nel 1476) di una studiatissima antologia che accoglie il fiore della produzione poetica italiana, dai Siciliani e Dante fino a Lorenzo il Magnifico: è la cosiddetta *Raccolta aragonese*, vero e proprio corpus di testi scelti inviato dal Magnifico a Federico d'Aragona figlio di Ferdinando re di Napoli, come repertorio di modelli letterari e linguistici.

Con la piena affermazione della stampa, e la connessa esigenza di migliore definizione del sistema della lingua, i nostri letterati tornano a consultare, con più precisa intenzione e con nascente scrupolo filologico, proprio i più antichi depositi della patria tradizione letteraria: a parte la curiosità di indagatori senza particolari intenti editoriali (come un Angelo Colocci che studia e postilla fittamente il *Canzoniere* vaticano e ne procura una copia), tutti i personaggi che nutrono propositi di ordinato studio della lingua sono specificamente impegnati a stabilire un legame «tra "testo" e "grammatica"»[3]; segna un momento cruciale della nostra storia linguistica la stretta concorrenza tra Pietro Bembo (che annuncia sue "notazioni della lingua" già nel 1500, allestisce le due aldine di Dante 1501 e Petrarca 1502, ed è a buon punto già nel 1512 nella stesura delle *Prose*, date alla luce in prima edizione nel 1525) e Gianfrancesco Fortunio (ideatore di un programma molto simile, maturo già nel 1509 e realizzato, limitatamente all'opera grammaticale, nel 1516), con i rispettivi editori e con altri autori al lavoro in quegli stessi anni (Calmeta, Claricio, Liburnio, Equicola, Alunno, Gabriele Trifone, Luna).

1.2 LA LESSICOGRAFIA DELLA CRUSCA. Richiamo appena il fatto che tutta la successiva, e più matura, grammaticografia e lessicografia italiana è obbligata ovviamente ad esibire i riferimenti testuali, che qualche autore di grammatiche esplicitamente ricompone in canone[4]. Nel campo della lessicografia, dopo un cenno a Francesco Alunno, che costruisce sul lessico delle "tre corone" e di altri autori *Le ricchezze della lingua volgare*, 1543, e *La fabrica del mondo*, 1546-48, i due principali strumenti lessicografici cinquecenteschi, il mio discorso deve soffermarsi sul *Vocabolario degli Accademici della Crusca.*

Quando gli Accademici (definitivamente costituitisi nel 1583) cominciano a lavorare a quest'opera, il canone di riferimento ormai vulgato per le descrizioni della lingua italiana è ancora fortemente limitato a testi strettamente letterari, toscani e trecenteschi (nella sistemazione bembiana, com'è noto, resta escluso Dante): seguendo Leonardo Salviati, i compilatori del *Vocabolario* superano d'un balzo questi limiti, si spingono ad autori quattrocenteschi e cinquecenteschi, anche non toscani, ed includono poi anche testi pratici e tecnici. La loro raccolta di "citati" raggiunge, fin dalla prima edizione (1612), il numero considerevole di 208 autori, con 309 opere (27 sono gli autori "moderni", con 52 opere). Per avviare un confronto tra il corpus di testi della prima Crusca e le dimensioni dei moderni corpora, è utile segnalare che nelle 1092 pagine *in folio* di questa edizione si contano 25.056 lemmi, articolati in varie accezioni, e 52.862 citazioni, per un totale di 1.152.999 parole (alle quali si aggiungono 391.816 parole della metalingua dei compilatori)[5].

---

[3] Cfr. Tavosanis 2001, pp. 55-76, a p. 56. Nei medesimi *Atti del Convegno di Gargnano del Garda* vedi anche Rabitti 2001 e Bologna 2001.

[4] Così fanno, ad esempio, Giacomo Pergamini, sia nel *Memoriale della lingua italiana* (1601) sia nel *Trattato della Lingua* (1613), e Daniello Bartoli, *Il torto e il diritto del non si può* (1655). Cfr. Robustelli 2006, rispettivamente alle pp. 102 e 119 sgg., e 286-289.

[5] L'intero contenuto delle quattro edizioni del *Vocabolario degli Accademici della Crusca* può essere consultato, in edizione digitale, attraverso il sito dell'Accademia della Crusca.

Sulla crescita delle dimensioni del corpus di testi nelle tre successive edizioni complete (1623, ancora in un solo volume; 1691, in tre volumi; 1729-1738, in sei volumi) offro dati soltanto per il numero degli autori e delle opere, rispettivamente: 246 e 372 (II. ed.); 342 e 614 (III. ed.); 383 e 659 (IV. ed.). Va anche segnalato che l'accrescimento delle citazioni riguarda in modo diseguale i lemmi: quelli di maggiore stabilità semantica vedono accrescere di poco, da un'edizione all'altra, il numero delle citazioni, mentre quelli che hanno potuto ricevere un nuovo carico semantico nell'uso della lingua vedono crescere di molto le citazioni. Ad esempio, per il lemma *ago* si passa da 9 citazioni nella I. edizione ad 11 nella IV., mentre per il lemma *corona* si passa da 9 a 27, e per il lemma *pane* da 12 a 64.

Non seguirò passo passo l'evoluzione della tradizione lessicografica italiana dopo il secolo XVIII. Lasciando da parte le vicende della quinta edizione della Crusca (intrapresa ed interrotta più volte nell'Ottocento; riavviata decisamente dopo la metà del secolo, interrotta definitivamente alla fine della lettera *O* nel 1923), basta constatare che la maggiore opera lessicografica ottocentesca portata a compimento, quella del Tommaseo e collaboratori, è ancora largamente, ma non completamente, basata su un corpus di testi e che solo l'impresa manzoniana del *Novo vocabolario*, dato il principio della ricerca dell'"uso" (fiorentino, contemporaneo) oblitera l'antica tradizione delle citazioni d'autore, adombrando il riferimento ad un "corpus di parlato".

2. CONCLUSIONI. Riprendo, in conclusione, l'accenno iniziale alla particolare utilità dei corpora testuali per la definizione dei problemi che pone l'uso odierno della nostra lingua. Ai non pochi dubbi sulla norma tuttora esistenti per noi parlanti e scriventi di oggi (vedi le alternative tra *gli* e *loro* pronome personale dativo plurale; alcuni usi dell'indicativo per congiuntivo; ecc.) potremmo dare risposte meno soggettive o cautamente sfumate, se dalla consultazione di corpora di grande ampiezza e costruiti secondo una adeguata tipologia dei testi potessimo ricavare che un determinato uso messo in discussione:
(a) è documentato con una certa stabilità nel corso degli ultimi duecento anni (in pratica, dall'incipiente rinnovamento della lingua scritta alle soglie dell'età romantica);
(b) è stato accolto, in questo arco di tempo, da un certo numero di autori di riconosciuta grande autorità, al di fuori di scelte stilistiche volutamente caratterizzanti sul piano diatopico e diafasico;
(c) è presente in una fascia di testi che fanno da ponte tra il parlato e lo scritto.

BIBLIOGRAFIA.

AVALLE
1992 *Concordanze della lingua poetica italiana delle origini (CLPIO)*, vol. I., a cura di D'Arco Silvio Avalle con il concorso dell'Accademia della Crusca, Milano-Napoli, Riccardo Ricciardi, MCMXCII.

BARBERA - CORINO - ONESTI
¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

BOLOGNA
2001 Corrado Bologna, *Bembo e i poeti italiani del Duecento*, in MORGANA - PIOTTI - PRADA 2001, pp. 95-122

LEONARDI
2000-01 *I Canzonieri della lirica italiana delle origini*, a cura di Lino Leonardi: vol. **I.**, *Il canzoniere Vaticano (Vat. lat. 3793). Riproduzione fotografica*; vol. **II.**, *Il canzoniere Laurenziano (Firenze, Biblioteca Medicea Laurenziana, Redi 9). Riproduzione fo-*

tografica; vol. **III.**, *Il canzoniere Palatino (Firenze, Biblioteca Nazionale Centrale, Banco Rari 217, ex Palatino 418). Riproduzione fotografica*; vol. **IV.**, *Studi critici*, Firenze, SISMEL - Edizioni del Galluzzo, 2000 (voll. I-III) e 2001 (vol. IV).

MORGANA - PIOTTI - PRADA
2001 Prose della volgar lingua *di Pietro Bembo. (Atti del Convegno di Gargnano del Garda, 4-7 ottobre 2000)*, a cura di Silvia Morgana, Mario Piotti, Massimo Prada, Milano, Cisalpino - Istituto Editoriale Universitario, 2001.

RABITTI
2001 Giovanna Rabitti, *Tra Bembo e Fortunio: una generazione inquieta*, in MORGANA - PIOTTI - PRADA 2001, pp. 77-94

ROBUSTELLI
2006 Cecilia Robustelli, *Grammatici italiani del Cinque e del Seicento. Vie d'accesso ai testi*, Modena, Mucchi Editore, 2006.

TAVOSANIS
2001 Mirko Tavosanis, *Le fonti grammaticali delle* Prose, in MORGANA - PIOTTI - PRADA 2001, pp. 55-76.

SITI DI RIFERIMENTO.

Crusca `http://www.accademiadellacrusca.it`

MARCO RICOLFI
(Università di Torino)

iij. Il terribile diritto.
La proprietà intellettuale: un incentivo od un ostacolo all'innovazione ed alla creatività?

> MEPH. *Es erben sich Gesetz' und Rechte*
> *wie eine ew'ge Narrheit fort;*
> Wolfgang Goethe, *Faust*, I.4, vv. 1972-1973.

0. LA QUESTIONE. Quando la professoressa Carla Marello ci ha chiesto di studiare quali fossero gli ostacoli di ordine giuridico che si frappongono all'accesso ed alla manipolazione dei dati linguistici che sono la materia prima per le operazioni di linguistica computazionale e quali sono i modi per poi fare circolare nel modo più ampio i risultati delle operazioni di "tokenizzazione", "markup" e di "tagging", la nostra sensazione è stata quella di essere chiamati ad una sfida difficile ma importante.

Difficile, perché si trattava di comprendere i dati base di una questione molto tecnica, che è il pane quotidiano per gli studiosi di un'altra disciplina, ma che subito si è rivelata di non facile mappatura per chi, come noi giuristi, non possegga i fondamenti di questo settore. Ma importante, anche, perché, armati della necessaria umiltà di chi sa di non sapere, abbiamo compreso che si trattava di un'altra frontiera di una battaglia in corso, quella per riportare la proprietà intellettuale alla sua funzione di incentivo all'innovazione ed alla diffusione della creatività, in un'epoca nella quale invece il diritto d'autore e gli altri diritti di proprietà intellettuale rischiano di essere utilizzati per bloccare l'innovazione invece che per favorirla.

0.1 UN POCO DI STORIA. Se poi siamo riusciti o meno a muovere qualche passo nella direzione giusta, lo dirà il lettore e soprattutto la comunità scientifica degli utenti. Per parte mia, vorrei far qualche passo indietro per ricordare che non sempre il diritto della proprietà intellettuale è stato il terribile diritto che troppo spesso oggi incontriamo nei crocevia presidiati *manu militari* dagli interessi di poche imprese dotate di potere di mercato.

La tutela dei brevetti per invenzione e del diritto d'autore è nata nel Settecento; ed è nata per aprire e non per chiudere. Il diritto d'autore è nato nel 1711 nell'Inghilterra della buona regina Anna. Esso ha consentito agli autori di affrancarsi dalla dipendenza dei mecenati e del potere politico. Forse l'atto di nascita dell'opinione pubblica e della libertà di espressione, nell'accezione moderna della parola, va ritrovato nell'anno 1775, quando il Dr. Johnson poté sferrare un poderoso e veemente attacco all'Earl di Chesterfield, che, in passato, era stato suo patrono. Il padre delle lettere inglesi poteva ormai tranquillamente contare sui proventi della sua infaticabile attività pubblicistica; e questi proventi gli erano garantiti dal *copyright*. Le invenzioni della rivoluzione industriale e dell'Ottocento si spiegano anche con l'incentivo fornito dal brevetto: senza il quale non è detto che avremmo avuto la rapida successione delle scoperte di Edison, di Bell. Perché non sempre gli individui e le imprese sono disposti ad investire in ricerca e sviluppo, se gli avanzamenti tecnologici possono essere immediatamente imitati dai concorrenti.

La proprietà intellettuale è però come un Giano bifronte. Conferisce un monopolio ai creatori; ed in questo modo fornisce un incentivo perché essi creino e perché trovino imprese disposte ad investire nello sfruttamento delle loro creazioni: l'editore che pubblica il *Dizionario* di Johnson e l'ATT che sfrutta l'invenzione del telegrafo. Ma lo stesso monopolio frena poi l'innova-

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. xvij-xix.

zione a valle: opere e invenzioni non possono essere imitate e neppur modificate se non con il consenso dei loro titolari. Fino ad un certo punto l'incentivo alla creazione è prevalso sul blocco dell'innovazione. Ma a partire da un certo punto, negli ultimi decenni del secolo scorso, è avvenuto il contrario.

Perché ad un certo momento il vento è cambiato? Le ragioni sono tante. Ma si possono ragionevolmente riassumere in una sola constatazione. In passato, il monopolio conferito da diritto d'autore e brevetti concerneva soprattutto beni materiali: i libri e poi i dischi; le invenzioni della meccanica prima, quelle della chimica e della farmacia dopo. Ora, il monopolio ha soprattutto per oggetto l'informazione. Soprattutto l'informazione in forma digitale e biotecnologica. Quello che possiamo dire con certezza è che se in passato i diritti di proprietà intellettuale rappresentavano delle isole di monopolio in un mare di concorrenza, oggi è vero il contrario: isole di concorrenza affiorano nel vasto mare dei diritti di monopolio.

1. COME RIAPRIRE? Il tema che quindi si è proposto oggi è quello di riaprire gli spazi di libertà che si sono chiusi nel frattempo; e di pensare ad una struttura dei diritti di proprietà intellettuale che ritorni ad essere capace di istituire un bilanciamento fra gli incentivi "primari" alla creazione e le istanze "secondarie" – ma non meno importanti – della disseminazione.

1.1 LE ISTANZE DELLA DISSEMINAZIONE. Queste "secondarie" istanze della disseminazione si stanno moltiplicando. Tutti richiamano l'esempio del software libero, che mette a disposizione i sorgenti in modo da consentire ad intere comunità, anzi a chiunque voglia, di sviluppare ed integrare i programmi di partenza in funzionalità sempre più ampie e nuove. Ma le istanze di libertà si moltiplicano. È possibile, oggi, digitalizzare biblioteche ed archivi; e l'Unione europea ci ricorda, con la sua iniziativa sulle Biblioteche digitali, che sarebbe un peccato se le opportunità di conservazione e di allargamento dell'accesso di testi, immagini, film, musiche, offerte dal digitale, venissero azzerate da una difesa ortodossa dei diritti d'autore. Esistono enti pubblici, come la BBC (ma anche la RAI) che hanno per decenni speso denaro del contribuente, derivante dal canone, per finanziare la produzione di programmazione *in house* (i telegiornali, le opere autoprodotte) o commissionarne la produzione. Perché ora il diritto d'autore dovrebbe sequestrarle in scantinati digitali?

Visto che la realtà di oggi è diventata tanto più complessa, è venuto il momento di pensare ad istituti della proprietà intellettuale più flessibili e più aperti. Uno di questi può consistere nell'adozione di licenze di Creative Commons, che sono basate sull'idea che non sempre gli autori di un opera nuova vogliano riservarsi tutti i diritti. I creatori che adottano CC si riservano solo alcuni diritti; gli altri sono consegnati a tutta la comunità, che può utilizzarli per fruirne, per creare opere nuove, per inserirli come tessera in nuovi mosaici. Come, per l'appunto, i corpora.

1.2 *ADELANTE, PEDRO, CUM JUICIO.* Vorrei dire – forse soprattutto come allievo di un grande maestro, Norberto Bobbio, amante delle aperture ma diffidente delle rivoluzioni – che questa esigenza di "ri"-aprire il diritto della proprietà intellettuale non intende mettere in discussione i tre assiomi, secondo cui la modalità di produzione dominante è costituita dal mercato, che essa contribuisce all'allocazione ottimale delle risorse e che la proprietà intellettuale continua ancor oggi a costituire un meccanismo prezioso per incoraggiare la creazione e lo sfruttamento primario di quel bene pubblico che sono le creazioni.

Per questo penso anche che, qualunque cosa si proponga, bisogna far attenzione a non uccidere la gallina dalle uova d'oro. Concretamente: se si consentisse alle biblioteche di digitalizzare i romanzi di Harry Potter, e metterli a disposizione online il giorno dopo che sono comparsi in libreria, si potrebbe stare sicuri che l'autrice non si prenderebbe la pena di scriverli e l'editore di stamparli. Quindi, l'Unione europea si preoccupa soprattutto di aprire alla digitalizzazione di opere fuori stampa, o delle opere cosiddette orfane, quelle che farebbero la loro figura come tes-

sera di un mosaico, salvo che non si sa a chi appartengano o come trovare i loro creatori. E certo non pensa all'accesso digitale ad Harry Potter.

È questa la ragione per cui nel nostro lavoro ci siamo posti anche una domanda non scontata. Va bene operare per allargare l'accesso ai testi da cui nascono i corpora; e va ancor meglio garantire l'accesso libero ai corpora stessi per fini di studio, di ricerca, di avanzamento culturale. E però: quale è allo stesso tempo la protezione di cui dispongono i medesimi corpora, intesi come output basato su dati linguistici altrui e dotato di un proprio valore aggiunto organizzato? La nostra risposta è che la tutela c'è; ed è data soprattutto dal diritto sulle banche dati. Cosicché alla libertà di certe modalità di utilizzazione può anche corrispondere un monopolio su certe altre; che va riconosciuto e difeso, perché può portare risorse economiche all'Università ed alle istituzioni che fanno ricerca. Se, dopo tutto, siamo alla ricerca di un nuovo equilibrio nel diritto della proprietà intellettuale, non è detto che questo debba soltanto portare a maggior accesso e maggior disseminazione. Non si può affatto escludere che questo porti anche a nuovi incentivi alla creazione ed all'utilizzazione "primaria", particolarmente benvenuti poi se i risultati economici vanno a vantaggio di istituzioni che fanno ricerca e continuano a fare ricerca di alto livello, nonostante tutto e nonostante tutti.

2. QUASI UNA CONCLUSIONE. Quest'ultimo accenno mi porta ad una riflessione finale. Che concerne il carattere inconsueto del nostro incontro, voglio dire: fra linguisti computazionali e giuristi. Che, però, a ben pensarci, forse non è poi così inconsueto per la città-laboratorio di Torino e per le tradizioni di ricerca torinesi. Che si propongono di aprire vie ed esplorare strade nuove; come è nella loro tradizione, che continua e che oggi si arricchisce di una nuova ambizione: di portare questa tradizione del nuovo in dote alla grande area metropolitana Milano-Torino che si sta profilando, e che non potrà non avere fra i suoi pilastri la riflessione teorica *als strenge Wissenschaft.*

MANUEL BARBERA
(Università di Torino)

iiij. La resa dei forestierismi in italiano. *Breve nota ortografica.*

0 PREMESSA. La strategia da adottare per i forestierismi, in un volume di questo genere, è argomento che va affrontato preliminarmente, non fosse che per giustificare l'editing che si è fatto del testo[1]. Il riferimento, va sùbito detto, è a quel tipo di antipurismo pragmatico e moderato che si ispira idealmente al Leopardi. V'è un passo dello *Zibaldone* che giova rileggere, idealmente sostituendo al francese l'inglese, ed alla lingua filosofica quella scientifica:

> *«Per li nostri pedanti il prendere noi dal francese o dallo spagnuolo voci o frasi utili e necessarie, non è giustificato dall'esempio de' latini* classici *che altrettanto faceano dal greco, come Cicerone massimamente e Lucrezio, né dall'autorità di questi due e di Orazio nella Poetica, che espressamente difendono e lodano il farlo.* [...] *Ben è vero che la greca letteratura e* [3193] *filosofia fu, non sorella, ma propria madre della letteratura e filosofia latina. Altrettanto però deve accadere alla filosofia italiana, e a quelle parti dell'italiana letteratura che dalla filosofia devono dipendere e da essa attingere, per rispetto alla letteratura e filosofia francese. La quale dev'esser madre della nostra, perocché noi non l'abbiamo del proprio, stante la singolare inerzia d'Italia nel secolo in che le altre nazioni d'Europa sono state e sono più attive che in alcun'altra. E voler creare di nuovo e di pianta la filosofia, e quella parte di letteratura che affatto ci manca (ch'è la letteratura propriamente moderna);* [...] *sarebbe cosa, non solo inutile, ma stolta e dannosa, mettersi a bella posta lunghissimo tratto addietro degli* [3194] *altri in una medesima carriera, volersi collocare sul luogo delle mosse quando gli altri sono già corsi tanto spazio verso la meta, ricominciare quello che gli altri stanno perfezionando; e sarebbe anche possibile, perché né i nazionali né i forestieri c'intenderebbono se volessimo trattare in modo affatto nuovo le cose a tutte già note e familiari, e noi non ci cureremmo di noi stessi, e lasceremmo l'opera, vedendo nelle nostre mani bambina e schizzata, qualla che nelle altrui è universalmente matura e colorita; e questo vano rinnovamento piuttosto ritarderebbe e impaccerebbe di quel che accelerasse e favorisse gli avanzamenti della filosofia, e letteratura moderna filosofica.* [...] *se vuol dunque l'Italia avere una filosofia ed una letteratura moderna filosofica, le quali finora non ebbe mai, le conviene di fuori pigliarle, non crearle da se* [sic]*; e di fuori pigliandole, le verranno principalmente dalla Francia (ond'elle si sono sparse anche nelle altre nazioni* [...]*), e vestite di modi, forme, frasi e parole francesi (da tutta l'Europa universalmente accettate, e da buon tempo usate): dalla Francia, dico, le verrà la filosofia e la moderna letteratura, come altrove ho ragionato; e volendole ricevere, nol potrà altrimenti che ricevendo altresì assai parole e frasi di là, ad esse intimamente e indivisibilmente spettanti e fatte proprie;* [3196] *siccome appunto convenne fare ai latini delle voci e frasi greche ricevendo la greca letteratura e filosofia; e il fecero senza esitare.* [...].*»*
>
> Giacomo Leopardi, *Zibaldone*, pp. 3193-6 = ed. Pacella 1991, pp. 1675-7

In pratica ciò equivale ad una certa generosità ad ammettere l'uso di termini di origine straniera ritenuti tecnicamente "indispensabili", ed una accettazione del loro ingresso, almeno iniziale, nella lingua come prestiti non adattati. Le ragioni ed i limiti di questa strategia sono stati diffusamente argomentati in Barbera - Marello a proposito della *corpus linguistics* (in corsivo; o meglio linguistica dei corpora in tondo), cui rinvio in tutto e per tutto, limitandomi qui a riassumerne le conseguenze pratiche contingenti, cioè ad enunciare le norme editoriali che si sono a questo proposito adottate in questo volume.

[1] Altre osservazioni puntuali saranno naturalmente fatte dagli autori nei loro contributi.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. xxj-xxij.

1. Il trattamento dei prestiti non adattati. Sinteticamente queste sono le norme che si sono seguite:

(a) vanno in tondo e non in corsivo in quanto parole non più straniere
(quindi: “file” e “corpus”, e non “*file*” e “*corpus*”).
(b) quanto alla formazione del plurale,
(1) i prestiti da lingue moderne rimangono invariati
(quindi: “i file” e non “i *files*”)
(2) i prestiti da lingue classiche sono pluralizzati come da grammatica (quindi: “i corpora” e non “i corpus”[2])
(c) la derivazione avviene secondo le normali regole italiane: prestiti non adattati in derivazione producono prestiti adattati
(quindi: “tag” >“taggare” > “taggato”)
(d) la ortografia originale viene tendenzialmente mantenuta in quanto distintiva anche delle famiglie derivazionali
(quindi: *token* > “tokenizzato”)
(e) le forme con trattino o spazio nell’originale se possibile sono univerbate con caduta del trattino o dello spazio[3]
(quindi: *mark-up* e *home page* > “markup” e “homepage”).

Bibliografia.

Barbera - Marello
2003 *i.s.* Manuel Barbera - Carla Marello, *Corpo a corpo con l’inglese della* corpus linguistics*, anzi, della linguistica dei corpora*, in *Atti del Convegno Internazionale Lingua italiana e scienze, Firenze, Accademia della Crusca 6-8 febbraio 2003*, in corso di stampa.

Pacella
1991 Giacomo Leopardi, *Zibaldone di pensieri*, edizione critica e annotata a cura di Giuseppe Pacella, Milano, Garzanti, 1991 “I libri della spiga”.

Sampson
2004 Geoffrey Sampson, *Introduction* to Sampson - McCarthy 2004, pp. 1-8.

Sampson - McCarthy
2004 *Corpus Linguistics. Readings in a Widening Discipline*, edited by Geoffrey Sampson and Diana McCarthy, London - New York, Continuum, 2004.

[2] Sono ormai abbastanza diffusi anche plurali invariati, e spesso tale comportamento è stato accettato dai lessicografi (che registrano il plurale “corpus” anziché “corpora”); noi, in ciò probabilmente da conservatori, continuiamo però a volere accordare un diverso *status* al lascito culturale della tradizione grecoromana. In parte il problema si è posto anche in inglese, dove accanto a *corpora* è apparso anche *corpuses*, al cui proposito molto britannicamente commenta Sampson 2004, p. 1: «it is quite permissible to Anglicize the plural and write *corpuses* – some corpus linguists use that form: we prefer *corpora* because *corpuses* sounds like ‘corpses’».
[3] Caso diverso però è quello di POS-taggato ecc., in quanto POS è una sigla mantenuta come tale in maiuscolo.

PARTE

.I.

MANUEL BARBERA
(Università di Torino)

1. Per la storia di un gruppo di ricerca[1]. *Tra bmanuel.org e corpora.unito.it.*

> [‒◡◡]. *ναυσιφορήτοις — δ' ἀνδράσι πρώτα χάρις*
> *ἐς πλόον ἀρχομένοις πομ- — παῖον ἐλθεῖν οὖρον· ἐοικότα γάρ*
> *κἀν τελευτᾷ φερτέρου νόσ- — του τυχεῖν.* [◡◡◡⏓]
> Pindaro, *Pitica* I., Ep. 2, stichi 1-3.

0. PREMESSA IN CIELO. «Era una notte buia e tempestosa» ... Padova, 14 marzo 1998, per volontà o (boulezianamente) per caso, due torinesi si incontrano in terra straniera. L'una, una lessicografa e linguista testuale, l'altro un filologo e linguista storico. Che, trasportati da insolita passione, si dissero: "facciamo della *corpus linguistics*!". Strani casi della vita.

Ed avrete certo capito chi erano i nostri due torinesi in incognito.

Di fatti, fu una vera *Kehre* nelle nostre carriere scientifiche e l'alba di una nuova e più ricca stagione di ricerca[2].

1. L'INIZIO DELLA RICERCA. La prima fase delle nostre ricerche si svolse prevalentemente (e contestualmente rispetto all'innesco padovano di ItalAnt) all'ombra della progettazione ed implementazione del CT o Corpus Taurinense (un corpus di italiano antico POS-taggato conformemente agli standard europei EAGLES/ISLE correnti per le lingue moderne, cfr. *infra* § 2.2.1). Ma fin da sùbito l'organizzazione dei lavori e la pianificazione delle nostre attività guardavano più lontano, a far nascere dall'esperienza del CT un gruppo di ricerca che si proponesse ambiziosamente di diventare uno dei più importanti centri italiani di creazione e diffusione di corpora.

Una preliminare ispezione delle risorse di *corpus linguistics* disponibili online (risultata anche nella costruzione della *CLR Guide*, una ricca guida annotata alle risorse di linguistica dei corpora e computazionale disponibili sul web) denunciava, infatti, assai chiaramente una grande esigenza di corpora liberamente disponibili, specie per la lingua italiana.

Collaborazioni importanti, frattanto, venivano avviate. Al di là di una breve parentesi con la commerciale Dima Logic, e di quella, assai proficua ma limitata al CT, con l'Opera del Vocabolario Italiano (OVI), la più importante è stata quella con l'Institut für maschinelle Sprachverarbeitung - Stuttgart (IMS), ed in particolare con il gruppo di ricerca di Ulrich Heid. La relazione con Stoccarda, tra le più felici, dura tutt'ora, ed ha apportato elementi fondamentali, primo tra tutte il Corpus WorkBench (CWB), inestimabile supporto informatico per tutti i nostri corpora, con il potente strumento di ricerca CQP (sul quale cfr. Christ - Schulze 1996, Christ et alii 1999 e qui Heid ¶ 4, *infra*) ed il POS-tagger Tree Tagger (cfr. Schmid 1994).

Frattanto veniva anche creata *bmanuel.org* (cfr. Barbera 2004, p. 126), una libera associazione privata di linguisti, filologi ed ingegneri, fondata e guidata da Manuel Barbera, attiva nel

[1] Nell'àmbito della giornata di studi, la comunicazione (semplice presentazione su Power point) di cui questo articolo è sommaria rielaborazione era più prolissamente intitolata *Come è nato e cresciuto www.corpora.unito.it: i corpora NUNC, Athenaeum, Valico, Vinca ed il corpus Taurinense.*

[2] La data è peraltro importante per la linguistica italiana in generale non fosse che per la fondazione del progetto ItalAnt (cfr. Renzi 1998), che è poi anche l'occasione per la quale avvenne l'incontro in questione tra Carla Marello e Manuel Barbera.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 3-20.

settore della linguistica dei corpora e dell'informatica umanistica, entro la quale svolgere tutte quelle attività di raccolta dati, preparazione di strumenti informatici appropriati, e confezione di corpora per le quali le nostre tradizionali strutture universitarie risultavano variamente inadeguate; la associazione è appoggiata ad un sito, in *housing* presso la pavese Sesamo (l'importante società fornitrice di servizi informatici globali), che diventa presto uno dei siti di linguistica più visitati a livello mondiale, oggi con una media di circa 5.000 accessi mensili.

Sul fronte universitario, invece, i nostri sforzi risultarono nella fondazione del Dottorato in Linguistica, Linguistica applicata, Ingegneria linguistica (attivo dal XVII ciclo)[3], necessario serbatoio formativo per un gruppo di ricerca.

L'ossigeno per queste ricerche fu poi trovato dapprima con un paio di progetti cofinanziati (COFIN), ma poi soprattutto con i Fondi per la ricerca di base (FIRB), che tanto si sono resi indispensabili per le ricerche di cui qui presentiamo i frutti.

2. LA PIENA DELLA RICERCA. Esaurite queste attività "fondanti" e creato un primo ed affiatato gruppo[4] di ricerca che, attorno a Carla Marello e Manuel Barbera, comprendeva già Marco Tomatis, Adriano Allora e Luca Valle (ed altri si aggiungeranno presto a loro), individuammo alcune generali linee maestre per le nostre ricerche (cfr. § 2.1), lanciammo diversi progetti di corpora, molti dei quali già liberamente consultabili (cfr. § 2.2), e creammo per questi un centro di distribuzione (cfr. § 2.4).

2.1 GLI INDIRIZZI. I principi guida delle nostre attività, come accennavo, si indirizzarono sùbito, quasi spontaneamente una volta presa consapevolezza attraverso la preparazione della CLR Guide dello status della disciplina e della sua accessibilità, lungo alcune linee ben precise, tra loro strettamente legate.

(a) I corpora creati dovevano essere di libera accessibilità (e ciò valeva soprattutto per l'italiano dove l'esistente[5] era spesso di dubbia legalità e/o con forti restrizioni d'uso), ed il loro strumento più efficace di diffusione era la loro consultabilità online.

(b) I corpora creati dovevano essere adatti anche per ricerche di tipo testuale (non fosse che per la nostra radicata tradizione di linguistica testuale), quindi interrogabili senza alcuna restrizione di contesto.

(c) Il problema degli aspetti legali dei corpora (acquisizione e licenze; l'esempio di GNU[6] nel software, ecc.) diveniva così centrale: le riflessioni che andavamo facendo su ciò sono qui riassunte in Allora - Barbera ¶ 5; la definizione legale della situazione è in Zanni ¶ 6; le "soluzioni" che abbiamo trovato sono presentate in Ciurcina - Ricolfi ¶ 7; e Ricolfi ¶ iij tratteggia infine l'orizzonte e la portata di questa operazione.

(d) Accanto all'accessibilità delle risorse ci appariva fondamentale la loro riutilizzabilità (cfr. Barbera 2001), in nome di una linguistica "ecologica", governata da una gestione responsabile ed "economica" delle risorse, preferibilmente gratuite e riciclabili.

(e) La necessità di giungere ad una definizione formale certa di cosa sia un corpus (indispensabile per la formulabilità stessa del "problema" legale di cui al punto c, ed i cui risultati si troveranno qui nei ¶¶ 5-7), da un lato, e dall'altra le esigenze di una ecologia delle risorse (di cui al punto d), hanno portato ad una riflessione su tokenizzazione, markup (caratteristiche

[3] Confluito nel 2006, col XXII ciclo, nella *Scuola di dottorato in Studi euro-asiatici: indologia, linguistica, onomastica*, Indirizzo in *Linguistica, linguistica applicata e ingegneria linguistica*.
[4] Ed anche il gruppo "tecnico" più ristretto, informalmente noto come "Le Tigri di Via Piazzi" (dalla ubicazione della sede dell'associazione), di cui sono membri fondatori Adriano Allora, Manuel Barbera e Marco Tomatis.
[5] La situazione oggi è un poco migliorata soprattutto grazie al forlivese Corpus "La Repubblica" di Marco Baroni (cfr. Baroni et alii 2004).
[6] Progetto «launched in 1984 to develop a complete UNIX-like operating system which is free software: the GNU system» (GNU homepage). Cfr. per maggiori dettagli qui oltre Allora - Barbera ¶ 5.

entrambe costitutive di un corpus: cfr. qui Barbera - Corino - Onesti ¶ 3, § 1) e sul POS-tagging (argomento che, a partire dall'esperienza del CT, diventerà anzi per noi fondamentale: cfr. *infra* § 3.1 e Barbera ¶ 8), ed in generale sulla standardizzazione dell'annotazione.

2.2 I RISULTATI: CORPORA. In questi otto anni quasi tutte le nostre energie, al di là della necessaria riflessione metodologica, metadisciplinare e programmatica, sono state convogliate soprattutto nel creare risorse, e ciò significa eminentemente corpora, anche a scapito di altri (a volte più gratificanti) aspetti della ricerca.

Complessivamente, abbiamo elaborato più di un miliardo e mezzo di token[7], e ne abbiamo già messo a disposizione online quasi mezzo miliardo; tra le lingue l'italiano, certo, costituisce il centro di questa produzione, ma non abbiamo trascurato neppure altre lingue. Una prima idea quantitativa del nostro operato al momento attuale è ricavabile dalla tavola seguente:

token elaborati	non taggati online	taggati online	taggati offline
IT	---	281.786.094	---
FR	---	315.260.061	---
EN	---	21.493.116	241.748.599
DE	---	---	304.533.385
ES	47.479.918	---	8.000
subtot.	47.479.918	364.831.411	546.289.984
tot. ***online***	412.311.329		---
tot. ***taggati***	---	911.121.395	
tot. ***elaborati***	958.601.313		

Tav. 1: totali dei token elaborati (primavera 2006) per etichettatura e disponibilità.

Più nel dettaglio, i corpora già online, da cui sono ricavate le cifre presentate nella Tav. 1, sono quelli riassunti nella tavola seguente (Tav. 2), dove sono fornite anche le specifiche di token, type e lemma.

I vari corpora, di cui nella tavola 2 sono presentati solo quelli di cui al momento esiste una versione disponibile già online, saranno poi descritti singolarmente (insieme ai loro compagni in corso d'opera o non ancora disponibili) nei sottoparagrafi seguenti.

[7] In altri termini, la nostra "produttività media" è stata di 119.825.164,125 token/anno, ossia ben 328.288,128 token/giorno: grossomodo un terzo di milione di parole al giorno per otto anni. Non crediamo di essere stati un cattivo investimento.

corpora online	subcorpora	token	type	lemma
Corpus Taurinense	---	259.299	21.087	7.599
VALICO	---	567.437	38.094	9.480
VINCA	---	64.652	9.323	3.859
Athenaeum Corpus	---	306.927	32.221	11.748
NUNC-IT	*general-I.*	127.708.505	1.346.652	42.531
	general-II.	109.692.794	1.098.829	42.252
	general-Tot	*237.401.299*	---	---
	cooking	4.161.627	187.544	23.543
	motor	7.909.608	273.744	23964
	photo	8.544.089	374.289	25.082
	cine	4.990.858	188.112	26.854
	photo-uncut	17.580.298	513.404	27.777
	NUNC-Tot	*280.587.779*	---	---
NUNC-ES	*general*	31.240.227	809.977	---
	cooking	2.098.489	118.250	---
	motor	13.415.613	487.288	---
	photo	725.389	30.956	---
NUNC-FR	*general-I.*	173.703.875	1.777.513	53.615
	general-II.	122.145.251	1.149.586	48.909
	general-Tot	295.849.126	---	---
	cooking	4.900.590	135.746	23.821
	motor	8.684.354	194.377	24.846
	photo	5.825.891	130.898	20.687
NUNC-UK	*motor*	12.426.186	226.654	38.773
	cooking	1.322.330	58.004	21.600
	photo	722.818	33.841	12.259
	business	7.021.782	146.691	39.112

Tav. 2: le cifre (token, type, lemma) dei corpora attualmente online (primavera 2006).

2.2.1 CORPUS TAURINENSE (CT). Come già detto, si tratta del corpus dal quale tutta questa avventura cominciò. Nella sua versione attuale comprende ventun testi fiorentini duecenteschi (da Brunetto, Bono, Rinuccino, Dante, Cavalcanti ed il Novellino, a testi mercantili, documentari e storici) per circa 250.000 parole, accuratamente tokenizzato, markuppato, POS-taggato e disambiguato[8] (cfr. Barbera - Marello 2000\03).

Vera punta di diamante della nostra produzione, la sua importanza nella nostra officina non è dovuta solo a ragioni affettive[9], ma anche ad altre più cogenti: (a) per i molteplici problemi posti dai testi antichi ha costituito una sfida ed una palestra tecnica ideale, contribuendo in modo formidabile alla formazione della nostra squadra; (b) di fatto rappresenta i risultati più

[8] Gli strumenti informatici usati per queste operazioni, quasi tutti in AWK, sono principalmente opera di Marco Tomatis, con minori contributi di Cesare Oitana, e minimi di Manuel Barbera. Per la disambiguazione in particolare cfr. più avanti in questo volume Tomatis ¶ 9.

[9] Particolarmente forti, peraltro, soprattutto in chi scrive, che è il solo filologo del gruppo.

perfezionati cui siamo per ora giunti nella preparazione di un corpus (in forza delle sue ridotte dimensioni unite al maggiore tempo - due bienni di COFIN - che abbiamo potuto dedicarvi); (c) è il primo privilegiato laboratorio su cui sperimentiamo ogni nuova tecnica, prima di esportarne le esperienze a nuovi corpora, trovandosi poi anche ad essere sempre il più *up-to-date*. Un esempio di ciò è il suo ruolo giocato nella nostra riflessione su tagset e POS-tagging (cfr. *infra* § 3.1 e qui Barbera ¶ 8).

2.2.2 ATHENAEUM CORPUS. Il nostro, particolare, omaggio al seicentenario della nostra università, in occasione del quale è stato reso disponibile online, vorrebbe documentare la produzione scritta di una grande Università italiana.

Si tratta, cioè, di un corpus di italiano scritto accademico, costruito con testi prodotti dall'Università di Torino, POS-taggati e classificato per argomento e tipo testuale. Le sue 3 componenti base, la cui preparazione è frutto del lavoro soprattutto di Luca Valle, (1) la rivista "L'Ateneo", (2) la *newsletter* "Dall'Università", (3) materiale amministrativo prodotti internamente o per il sito di ateneo, di cui la terza è ancora in implementazione, saranno presto interrogabili anche autonomamente online (già lo sono in locale). I due contributi in questo volume che si basano su di esso (Cignetti ¶ 11 e Ferrari - Mandelli ¶ 10) utilizzano infatti solo il primo subcorpus.

2.2.3 VALICO. Si tratta di un innovativo *learner corpus* di italiano scritto, il cui nome (allusivo sia dei monti del Piemonte sia del processo di apprendimento) è acronimo di Varietà di Apprendimento della Lingua Italiana: Corpus Online (cfr. Barbera - Marello 2004).

Nato nel 2003, e contributo al settore scientifico disciplinare cui M. Barbera e C. Marello càpitano di appartenere, questo corpus internazionale di apprendenti italiano è POS-taggato ed arricchito con un dovizioso markup testuale e sociolinguistico, che è stato recentemente migliorato nella sua organizzazione ed interrogabilità dai lavori di Schaupp 2006. Non ne diciamo oltre solo perché sarà l'oggetto di Corino - Marello 2007 e Corino - Heid - Schaupp 2007 *i.p*.

2.2.4 VINCA. Corpus di italiano scritto di nativi, nato nel 2004 come gemello a VALICO per fungergli da monitor (funzione cui il suo nome, acronimo di Varietà di Italiano di Nativi Corpus Appaiato, allude), ha presto sussunto anche dignità e vita autonoma.

La sua prima beta, da tempo disponibile in locale, è stata recentemente messa online.

2.2.5 NUNC. È forse il progetto più originale e strettamente legato al FIRB: si tratta di una collezione multilingue[10] di corpora di lingua contemporanea, tanto generici quanto specialistici[11], basati sui messaggi dei newsgroup; il nome, allusivo alle sue caratteristiche di *Umgangssprache*[12] contemporanea, è infatti acronimo di Newsgroups UseNet Corpora (cfr. oltre Corino

---

[10] Le lingue coperte dal progetto sono per ora danese, estone, finnico, francese, italiano, inglese britannico ed australiano, portoghese, spagnolo continentale e cileno, tedesco, ed ungherese.

[11] I settori specialistici su cui abbiamo per ora sperimentato sono quelli dell'alimentazione, della fotografia e dei motori, con escursioni anche al diritto ed al *business*, ma ovviamente in futuro se ne potranno studiare altri ancora.

[12] La nozione è vetusta, legata soprattutto alle problematiche sorte intorno al cosiddetto "latino volgare" tra i grandi *patres* della romanistica; già lo Spitzer, inoltre, in diversa ma confrontabile ottica, la aveva applicata all'italiano ("*italienische Umgangssprache*": cfr. Spitzer 1922/2007); e, comunque, è stata riproposta anche recentemente (cfr. Kiesler 2006). L'analogia sembra abbastanza buona, in quanto si tratta, molto in soldoni, di una lingua comune, usuale e media, non tematicamente o sociologicamente delimitabile, più vicina al parlato ma di fatto scritta, e per la quale, in realtà la dicotomia scritto-parlato non è veramente pertinente.

¶ 13 e Barbera 2007 *i.s.*). Nato nel 2002 per iniziativa di M. Barbera[13], che ne indovinò l'utilità ed iniziò i primi download sperimentali di testi nell'inverno 2001, fu da questi proposto come principale fonte testuale del progetto FIRB (cfr. Barbera 2004 *in.*); da allora vi hanno lavorato pressoché tutti i membri del nostro gruppo[14].

Un newsgroup è un forum telematico a libero accesso, gratuito, disponibile su Internet, che si manifesta nella forma di testi scritti, ed il cui funzionamento è assai semplice: ogni utente scrive un messaggio, il post, e lo invia ad una specie di "bacheca elettronica" mantenuta presso una rete di server (i newsserver che costituiscono UseNet), dai quali gli altri utenti del gruppo possono scaricarlo, leggerlo e rispondervi, costruendo anche articolate catene (thread) di botte e risposte. La facilità d'uso garantisce la grande diffusione dello strumento tra le categorie più diverse di utenti e giustifica la grande quantità di traffico esistente su UseNet. Queste "bacheche elettroniche" che sono i newsgroup sono poi articolate in una tassonomia precisa, ossia in un sistema di cornici argomentative che si chiamano "gerarchie"[15], a base geografico-nazionale e/o tematica; anche queste gerarchie, peraltro, nascono dal basso in base alla iniziativa degli utenti.

I vantaggi di questa base testuale per la *corpus linguistics* sono numerosi: (a) la grande abbondanza testuale; (b) il presentare una *Umgangssprache* assolutamente contemporanea e reale molto variata per registri e temi; (c) la presenza di gerarchie classificate tematicamente dal basso; (d) l'organizzazione in gerarchie nazionali che è garanzia di uniformità diacorica; (e) la verosimile disponibilità legale del materiale[16]; (f) l'interesse testuale del fenomeno del quoting; (g) l'interesse lessicografico, antropologico e sociologico dell'essere UseNet una sorta di "enciclopedia popolare", organizzata secondo una "folk taxonomy".

A fronte di questi, a mio parere irresistibili, vantaggi ed aspetti di interesse, il ricorso a UseNet presenta anche degli indubbi svantaggi, il cui peso complessivo è però assolutamente minore: (a) peculiarità linguistiche mediate dal mezzo (gergo informatico, abbreviazioni, emoticon, ecc.); (b) frequenti "sporcature" del testo dovute alla trasmissione (passaggio da charset diversi, ecc.) od alla battitura; (c) presenza di spam, post OT ("out of topic") e crossposting; (d) l'abbondanza di testo ripetuto, a volte (quando effetto del quoting) testualmente rilevante e quindi "buono", ma comunque sempre per statistiche lessicali dannoso.

Gli aspetti problematici evidenziati sono stati (anche se ancora non del tutto) ovviati da una complicata preparazione dei testi, attuata attraverso vari moduli di filtraggio, tokenizzazione e markuppatura[17].

---

[13] L'unico precedente importante in tal senso è ELWIS (cfr. Hinrichs et alii 1995 e Feldweg - Kibiger - Thielen 1995), di cui peraltro presi conoscenza solo successivamente. L'impresa di questo corpus è peraltro assai rilevante anche per lo sviluppo dei tagset (campo sul quale ci stiamo anche noi esercitando: cfr. *infra* § 3.1 e Barbera ¶ 8), in quanto presentò la prima proposta di tagset tedesco poi confluita nel STTS tagset (cfr. Schiller et alii 1999). Il *CMU Text Learning Group Data Archive* di Tom Mitchell del 1993, di solito noto come "20 Newsgroups", non può invece intendersi come un vero precedente, in quanto, secondo la definizione proposta in questo volume (Barbera - Corino - Onesti ¶ 3, § 4), non si tratta tanto di un corpus quanto di una collezione di testi allestita per test di *machine learning*.

[14] Tra le prime, sperimentali, ricerche condotte a partire dai NUNC ricordiamo Valle 2006 (ma 2004) e 2005 *i.s.*

[15] I loro nomi sono infatti costruiti gerarchicamente, ad esempio `it.diritto`, `it.diritto.condominio`, `it.diritto.assicurazioni`, ...; `it.discussioni.animali`, `it.discussioni.animali.gatti`, `it.discussioni.animali.cani`, `it.discussioni.auto`, `it.discussioni.auto.ford`, …; ecc.

[16] UseNet per definizione e tradizione è il regno del pubblico dominio, quindi ciò sembrerebbe una ovvia assunzione; in realtà, se lo si dovesse sostenere legalmente, le cose potrebbero non essere così pacifiche (talvolta si è ricorso ad un cosiddetto "diritto implicito"), ma dato che il comune sentire sostiene comunque la nostra *bonam fidem*, e che non vi sono ad ogni buon conto interessi rilevanti lesi, è certo assai improbabile che contestazioni significative possano essere sollevate. In effetti sono anni che Google mantiene commercialmente archivi di newsgroup senza che ciò sia avvenuto.

[17] I "tools" fondamentali per queste operazioni sono stati approntati in Perl da Sara Casavecchia (cfr. Casavecchia 2005) e Simona Colombo; sono inoltre in corso una revisione della struttura dei metadata, in base ai lavori di Schaupp 2006 su VALICO, ed un approfondimento della marcatura dei confini di frase (cfr. Onesti *i.p.*).

Nonostante la quantità e dimensione dei corpora già preparati, solo una piccola parte dei materiali raccolti è stata finora elaborata, puntando soprattutto alla costruzione di corpora specialistici, e/o funzionalizzati alla ricerca lessicografica e terminologica (con abbattimento del testo ripetuto a scàpito dell'integrità dei thread), e solo più raramente a quella testuale (con mantenimento dell'integrità dei thread a scapito della presenza di molto testo ripetuto; cosa che ha reso possibile studi come Marello 2007). Molte vie sono ancora aperte, non ultimo quello della costituzione di una serie di *monitor corpora* (materiali in questo senso sono già stati scaricati per l'italiano ed il tedesco).

2.2.6 SMS. Si tratta di un *monitor corpus*, per ora di dimensioni assai modeste, di messaggini telefonici (il suo nome è infatti un acronimo quasi ricorsivo: *SMS Monitor Studies*), ideato e mantenuto dal 2003 da Adriano Allora, a conferma del nostro interesse in generale per i moderni linguaggi della comunicazione mediata (per cui in generale cfr. soprattuto Allora 2005 ed *i.p.*).

Propriamente è «una raccolta aperta di testi strutturati, nella fattispecie etichettati attraverso inserimento nel database, senza ambizioni di bilanciamento», dato che i testi sono immessi volontariamente dagli utenti medesimi; pertanto «esso rappresenta la varietà di italiano scritto per mezzo del telefono cellulare in un certo senso per accumulazione, nel suo divenire e trasformarsi» (homepage di e-allora.net).

2.2.7 JUS JURIUM. Il più giovane tra i nostri virgulti, nato nel febbraio 2005 per iniziativa di chi scrive[18], e quindi curato da Cristina Onesti e me, è un corpus in lingua italiana che intende coprire la totalità dell'universo di discorso legale oggi corrente in Italia[19]. In latino *jus jurium* vale 'minestrone di diritti': la molteplicità dei tipi di discorso legale, che il corpus vuole documentare, e la curiosa omofonia tra 'diritto' e 'minestrone' in latino hanno infatti ispirato il suo nome.

Di concezione innovativa, tanto da giustificare lo spazio che accordiamo qui alla sua presentazione, il corpus è etichettato per parti del discorso ed ha un robusto markup testuale e diplomatico. Tra le sue finalità, in particolare, vi è proprio quella di poter interrogare in modo "ricco" i testi, intersecando la loro definizione diplomatica con il loro assetto linguistico e testuale. Jus Jurium propriamente è un insieme di più subcorpora: attualmente, si sta lavorando a tre subcorpora, che seguono, per così dire *von Wiege zum Grabe*, tutta la "vita" delle leggi, dal loro concepimento nelle discussioni parlamentari, alla loro codificazione in regole normative, alla loro applicazione nei procedimenti giudiziari. In futuro speriamo di aggiungere altri due subcorpora, uno dedicato all'insegnamento della Legge, e l'altro a come le persone "comuni" parlano di solito di legge.

In sintesi, l'articolazione generale di Jus Jurium nella sua concezione più ampia sarà la seguente: 1. *Sectio Parlamentaris* consistente negli "stenografici" delle sessioni delle Camere e delle Commissioni camerali, dei vari Atti di indirizzo e di controllo e dei Disegni di legge; 2. *Sectio Normativa* consistenti nella Costituzione, nei Codici, nelle leggi e nei decreti di Governo, Ministri, Regioni ed Autorità amministrative autonome, e in una campionatura di testi paranormativi; 3. *Sectio Judicialis* consistente negli atti ("sentenze" ecc.) pronunciati dalle varie Corti, di tutti i gradi; 4. *Sectio Didactica* sarà implementata se otterremo il copyright di qualche

[18] Certo invogliato dalla disponibilità legale di tali materiali (in base all'art. 5 della legge 22 aprile 1941), ma soprattutto spronato dalle stimolanti ricerche di Mortara Garavelli 2001, ed assecondato dalla incomparabile cortesia e disponibilità di Mario Garavelli, che non potrei mai ringraziare abbastanza.

[19] Molti dei testi necessari al progetto erano liberamente raccoglibili in base alla legge 22 aprile 1941 n. 633 "Protezione del diritto d'autore e di altri diritti connessi al suo esercizio", il cui art. 5 stabilisce che «Le disposizioni di questa legge non si applicano ai testi degli atti ufficiali dello stato e delle amministrazioni pubbliche, sia italiane che straniere».

rappresentativo manuale di Diritto; 5. *Sectio Communis* sarà esportata dal NUNC italiano, selezionando i newsgroup di interesse legale. Buona parte della Sectio Normativa è ormai pronta, ed attende solo l'*encoding* in CQP.

Oltre alla dimensione relativa al tipo di discorso giuridico (riflessa nella organizzazione in sezioni del corpus), si è tenuto conto, quando possibile, anche della, trasversale, dimensione della estensione locale: le sezioni 2. e 3. infatti comprendono testi tanto nazionali quanto regionali; in un futuro potrà espandersi anche ai testi europei ed internazionali (purché in lingua italiana).

Quanto alla scelta dei testi, è da notare che la loro "rappresentatività" è assai particolare, soprattutto dal punto di vista della dimensione della cronologia. Il concetto di contemporaneità qui si configura infatti in modo più problematico dell'usuale, data la curiosa natura astorica dei testi legali: anche un Regio Decreto, se mai revocato, è tuttora in vigore; e facendo parte della normativa vigente è issofatto attuale e "contemporaneo". La "contemporaneità" dei materiali, oltre che *de jure*, è comunque garantita anche *de facto* dalla loro presenza online in più siti, anche di carattere non istituzionale, il che li garantisce come testi rappresentativi in quanto effettivamente presenti nell' "uso" attuale. Un bilanciamento[20], inoltre, verso la nozione ordinaria di contemporaneo, è stato comunque introdotto favorendo, quando possibile (ad es. per i testi parlamentari, non normativi, per cui non vale il discorso "legale" sopra accennato), i testi prodotti nell'ultima legislatura.

2.3 I RISULTATI: ALTRE RISORSE. Oltre i corpora medesimi, molte altre risorse di diverso tipo sono state prodotte in questi anni. Innanzitutto, oltre ai tools di preparazione dei vari corpora (notevoli, come accennato, sono soprattutto le batterie di programmi allestite per CT e NUNC), i risultati "software" più degni di menzione sono i seguenti.

E_NT_ER (ENgine for TExtual REsearchers), un motore per ricerche di linguistica testuale ideato e scritto da Adriano Allora intorno al progetto di VALICO, e recentemente presentato all'ultimo congresso SLI (Allora 2006 *i.s.*). Le sue due caratteristiche principali sono: (a) che è adatto quindi per lavorare con testi brevi corredati da molti metadata; e (b) che, sviluppato in Perl, si caratterizza per essere volutamente *developer-friendly*, ossia mantiene il codice in cui è programmato accessibile anche a non programmatori.

SMORFIA (SMOR[21] Finite states Italian Analizer), un analizzatore morfologico della lingua italiana, robusto e disponibile sotto GNU, sviluppato da Marco Tomatis come tesi dottorale (Tomatis 2004; cfr. Tomatis 2006 *i.s.*), capace di mostrare all'utente l'intera struttura fonomorfologica dei verbi italiani[22]. Il programma agisce come un normale strumento per l'analisi

---

[20] Anche il problema del bilanciamento secondo la dimensione della varietà testuale, inteso come la quantità dei testi da scaricare per ogni tipologia, era rilevante soprattutto per i subcorpora 2. e 3. Tale bilanciamento si può, infatti, intendere in due modi: (a) basato primariamente sulla loro reperibilità in rete, nell'idea che questa situazione di fatto rappresenti un ottimale "bilanciamento naturale"; (b) basato sulla rappresentatività ed importanza normativa (esiste una riconosciuta gerarchia di importanza delle fonti di normativa) o giurisprudenziale (l'uso nelle raccolte di giurisprudenza è in genere di privilegiare la Corte di Cassazione rispetto alle Corti di merito) degli atti. Nel caso del subcorpus normativo si è tentato una modesta correzione di (a) in base a (b), in quanto i due criteri avrebbero prodotto risultati non perfettamente collimanti. Nel caso del subcorpus giudiziario, invece, i due criteri hanno prodotto risultati quasi collimanti (con l'unico problema che a causa della scarsa reperibilità online dei testi, tuttavia, la presenza di alcune Corti "minori" è comunque troppo bassa per essere rappresentativa).

[21] SMOR (*Stuttgart Morphologie*) è la morfologia computazionale del tedesco sviluppata e compilata a Stoccarda usando SFST (*Stuttgart Finite State Transducer)*, ossia un «toolbox for the implementation of morphological analysers and other tools which are based on finite state transducer technology» secondo recita la sua homepage: cfr. Fitschen - Heid - Schmid 2004. SMORFIA ne verrebbe a formare il corrispondente italiano.

[22] Per ora il sistema si limita a questa particolare area della lingua, tuttavia l'implementazione di ulteriori parti del discorso è in attualmente in corso d'opera.

morfologica, ma con un approccio innovativo: accettando in ingresso sia interi documenti, sia singole parole introdotte tramite tastiera, e permettendo inoltre il ridirezionamento dell'uscita sia su schermo sia su un file specifico, comunica infatti all'utente non solo i diversi valori che l'elemento flessionale può di volta in volta assumere, ma anche (discostandosi dalla maggior parte dei progetti analoghi) la sua struttura completa, correttamente suddivisa nei suoi costituenti principali mostrati sequenzialmente all'interno della stessa stringa di testo come coppie attributo-valore. E per ottenere una struttura così elaborata, organizza le regole di analisi del trasduttore a stati finiti in una maniera tale da garantire un approccio di tipo *Item-and-Arrangement* in luogo del classico *Item-and-Process*.

MorFo (MORfemi FOndamentali), un analizzatore morfologico a scopi didattici preparato nel 2004 da Elisa Corino e Simona Colombo (cfr. Corino - Colombo 2004, Corino 2006 e Colombo 2006). Si tratta di un aiuto tanto per l'insegnante per fare esercitare la comprensione di testi, quanto per lo studente per memorizzare i morfemi nella lettura autonoma di testi: è pertanto uno strumento ibrido, parte glossario (come glossario alfabetico con più di 300 prefissi e suffissi derivativi e compositivi italiani, facilita la categorizzazione e la memorizzazione dei morfemi, nonché costituisce una base per lavorare sulla formazione delle parole, sulle loro funzioni e forme in testi) parte strumento didattico (in quanto mezzo per l'autoformazione del docente, aiuto nella selezione dei testi da far leggere, e suggeritore di domande di comprensione perché evidenzia i "grumi" difficili per forma e quindi, probabilmente, per contenuto).

ClitRec (A CLItic RECognizer), un software AWK creato nel 2004 da Marco Tomatis (cfr. Tomatis 2005) per il riconoscimento delle forme enclitiche presenti in un corpus tokenizzato, ma non ancora etichettato. Fondato su regole linguistiche che stabiliscono inferenze mediante l'analisi della capacità di ciascuna forma enclitica di selezionare una particolare tipologia verbale, il sistema richiede per il suo corretto funzionamento l'esistenza di un formario di macchina. Costruito per l'italiano, la sua architettura estremamente flessibile lo rende comunque applicabile a qualsiasi altra lingua, cambiando opportunamente regole e formario.

ILV$_A$T (Indice di Leggibilità per VArietà Testuali), un indicatore di leggibilità variabile preparato da Adriano Allora. «Nato in seno ad un progetto che si propone di indagare le varietà testuali, ad ogni livello – di genere testuale, mediale, stilistico –, è stato pensato per valutare la leggibilità di un testo sulla base di un file di parametri elaborato apposta per quel tipo di testo: lettera commerciale, articolo scientifico, bando di concorso, circolare amministrativa, racconto pornografico, verbale di riunione ...» come spiega Allora nella homepage; «Ogni file di parametri descrive, quindi, un indice di leggibilità specifico per un tipo di testo. Dunque quando si usa ILV$_A$T lo si può fare in diversi modi: è possibile creare un nuovo file di parametri adatto ai propri scopi [...]; è possibile usare un file di parametri esistente e disponibile nel database per valutare il proprio testo [...]; è possibile modificare un file di parametri esistente, segnalandolo come modificato [...]; è possibile aggiornare un file di parametri [...].» (*ibidem*).

Non bisogna, infine, neanche trascurare, tra i "risultati", l'organizzazione di importanti convegni (soprattutto quello Euralex 2006), coronata dalla pubblicazione degli Atti (Korzen - Marello 2000, Corino - Marello - Onesti 2006, ed il presente volume stesso); né la pubblicazione di un importante volume (Mortara Garavelli 2001) da parte della decana del nostro gruppo (che già diresse il primo COFIN sul Corpus Taurinense); né infine la didattologicamente rilevante (non fosse che per le istituzioni coinvolte) opera collettanea Bosc - Marello - Mosca 2006.

2.4 LA DISTRIBUZIONE. Accanto alla definizione di chiare linee direttive caratterizzanti (cfr. § 2.1), ed al conseguimento di rilevanti risultati (cfr. §§ 2.2 e 2.3), bisognava sùbito pensare a degli adeguati mezzi di diffusione e condivisione delle nostre ricerche. Lo strumento principe per ciò è stato presto individuato nella messa in rete con consultabilità online (senza

ovviamente escludere, a richiesta, la consultazione in locale o l'invio di specifici pacchetti per FTP o CD-ROM).

Si trattava, cioè, di disporre di server e banda con caratteristiche adatte per intensive query online.

Individuati i requisiti necessari ed i fondi disponibili, nell'autunno 2003 venne installato, grazie anche all'interessamento di Ferdinando d'Isep, un Compaq Server 380 (biprocessore Pentium IV con clock a 2.8 GHz, 2 GB di RAM e 6 hard disk in Raid 5) presso il Centro ReTe (Centro di interesse generale di Ateneo Reti e Telecomunicazione) dell'Università di Torino: era così nato corpora.unito.it.

3. PROGETTI IN CORSO E FUTURE INIZIATIVE. Quello fin qui presentato è solo il bilancio provvisorio, pur già cospicuo, di un'attività che è tuttora in pieno corso, e che non ha alcuna intenzione nel presente e nel futuro di arrestarsi ripiegando sui propri allori. In particolare, le questioni al momento più rilevanti e le agenda più impegnative per il futuro sono le seguenti.

3.1 PERFEZIONAMENTO E STANDARDIZZAZIONE DEI TAGSET. Dell'importanza del POS-tagging e della costruzione di tagset adeguati ci eravamo ben resi conto nella preparazione del CT, il cui tagset rappresenta a mio parere un risultato assai importante (cfr. qui Barbera ¶ 8). Col crescere delle lingue di cui ci stavamo occupando, aumentava anche il numero di tagset con cui ci dovevamo confrontare. Il TreeTagger, lo strumento principale di annotazione di cui ci siamo serviti, era dotato di file di parametri con lo *STTS tagset* per il tedesco (cfr. Schiller et alii 1999, e *supra* nota 13) e l'*EPADES tagset* per l'italiano, cui presto si aggiunsero il *PennTreebank tagset* per l'inglese (cfr. Santorini 1990/1 e Marcus - Santorini - Marcinkievicz 1994), un *EPADES-like tagset* per il francese e recentissimamente un *CRATER-like tagset* per lo spagnolo. Ed ovviamente, come primo punto di partenza, abbiamo usato quel che c'era a disposizione.

Tale molteplicità di tagset, di cui tra l'altro la maggior parte non gerarchicamente tipati (per le nozioni cfr. Barbera ¶ 8), anche se inevitabile nelle fasi iniziali, è lungi dall'essere ottimale. Stiamo infatti ora sperimentando vari approcci per un'unificazione almeno dell'architettura, a favore di una struttura tipata quale quella del CT-tagset. I lavori sono particolarmente avanzati per l'italiano moderno e soprattutto per lo spagnolo[23], di cui sono di imminente rilascio tagset e *parameter file* per il TreeTagger. Un mapping tra i tagset attualmente usati in bmanuel.org / corpora.unito.it è presentato oltre in Barbera ¶ 23.

Per dare un'idea (un'esemplificazione più dettagliata sarà presentata oltre in Heid ¶ 4) di cosa si può ottenere con l'applicazione di un tagset CT-like e l'uso di un motore di ricerca potente come il CQP, si considerino le due query seguenti nel CT:

```
[1]    [word = ".*e" & pos = ".v.+.+.ind.ipf.*" & kat = ".*2.+.6.*"]
[2]    [lemma="per"][lemma="che" & !pos=".conj.*" & !pos=".*.rel.*"]
```

La prima permette di cogliere tutti gli indicativi imperfetti di seconda persona singolare con terminazione anomala *-e*; la seconda consente di cogliere tutte le interrogative indirette introdotte da *per che*: il grado di raffinatezza di ricerca ottenuto si converrà che è assai elevato.

Accanto alla standardizzazione dei tagset, si è avviato anche un programma (Adriano Allora) per la costruzione di interfaccia di ricerca personalizzati, di cui alcuni risultati sono già visibili su corpora.unito.it

[23] A prevalente opera per il primo di Manuel Barbera e Marco Tomatis, e per il secondo di Manuel Barbera, Margarita Borreguero Zuloaga e Marco Tomatis. I lavori sullo spagnolo si sono giovati anche della tesi di laurea (2004-5) di Giovanna Brino. Il tagset, con pochi commenti, è già stato anticipato in Barbera 2007 *i.s.* e figura, nella versione corrente ed aggiornata (la 1.2), qui *infra* in Barbera ¶ 23.

3.2 PROSEGUIMENTO DI CORPORA AVVIATI. Di tutti i progetti di corpora presentati in § 2.2 e sottoparagrafi è prevista una continuazione.

In particolare, dei progetti in fase finale o comunque avanzata, per il CT è in cantiere una versione ampliata, di Atheaeum è previsto il completamento, e di VALICO e VINCA la progressiva crescita con nuovi materiali e formati di annotazione sempre più raffinati. Il cantiere di Jus Jurium è invece nelle sue fasi iniziali. Anche i lavori su NUNC, pur avendo già conseguito risultati significativi, devono essere considerati alle loro fasi iniziali, soprattutto considerandone le potenzialità di sviluppo: le più importanti sono l'allargamento dei corpora disponibili alle altre lingue, la costruzione di *monitor corpora* (prevista per l'italiano ed il tedesco) ed il miglioramento dei procedimenti di filtraggio e markuppatura.

3.3 NUOVI CORPORA. Accanto a queste attività già avviate è previsto (od appena iniziato) il varo di altre iniziative.

La maggior parte dei nuovi progetti riguardano l'italiano. Materiali sono stati acquisiti per corpora di italiano scritto giornalistici da *La Stampa* (Cronaca) e *La Valsusa*, per il quale ultimo sono già in corso anche le trascrizioni e markuppatura dei testi. In un panorama peraltro già abbastanza affollato, è infatti soprattutto il secondo a rivestire particolare interesse, essendo centrato non sulla stampa generica, ma su quella regionale e locale dei giornali a piccola diffusione (*La Valsusa*, appunto), le cui caratteristiche non sono state finora particolarmente indagate, pur rappresentando una categoria rilevante nel panorama giornalistico italiano. Accanto a ciò, sono stati avviati i primi contatti editoriali per la creazione di un Corpus di italiano scritto letterario, in previsione, innanzitutto, della preparazione di quel Grande Corpus Bilanciato di Italiano Scritto Contemporaneo di libero accesso, sogno della linguistica italiana, e che potremmo, in un futuro sperabilmente non troppo lontano, essere in grado di produrre.

Non mancano, inoltre, progetti minori e legati a situazioni specifiche. I più rilevanti sono un CORpus internazionale di etichette del VINO (CorVino), in avanzata fase di raccolta testi (cioè consumazione bottiglie!) e definizione di *Guidelines*, un corpus di testi di equitazione (EquUs "*EQUitationis corpUS*"), ed un corpus audio di cinese standard, la cui preparazione è da tempo quasi ultimata, ma la cui messa online è sempre stata rimandata da contrattempi vari.

4. E POI? Troppo poco? Fieno in cascina non ne difetta certo, e viene spontaneo domandarsi se mai ce la faremo. Siamo, infatti, un gruppo assai determinato e di grandi lavoratori, ma molto piccolo e con difficoltà di sostentamento non indifferenti.

Nei crudi noviluni annebbiati di questo secolo che vano sarebbe immaginare dei precedenti meno superbo e sciocco, non possiamo certo sperare in trionfi pitici come i naviganti in epigrafe; speriamo, però, almeno che ad un più inflazionato trofeo corrispondano meno impegnativi cimenti, e che pur sempre il vento propizio all'inizio della navigazione sia anche augure di un felice ritorno al porto: d'altronde, oggi come allora, il Sole, quel *busy old fool*, oltre che risplendere sulle sventure umane, sembra continuare a volere, ogni tanto, circonfondere *λαμπρὸν φέγγος* agli audaci.

Al più, quando saremo canuti e stanchi, se non ci sarà venuto un coccolone prima, potremo sempre fare un nuovo corpus:

CRAP: *Corpus di Recriminazioni di Accademici in (forse) Pensione*.

BIBLIOGRAFIA.

AA. VV.
2004 *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, ELDA, 2004.

ALLORA
2005 Adriano Allora, *A Tentative Typology of Net Mediated Communication*, comunicazione presentata alla *Corpus Linguistics 2005 Conference, Birmingham July 14-17 2005*, disponibile online alla pagina `http://www.corpus.bham.ac.uk/PCLC/`
2006 *i.s.* Adriano Allora, *EnTeR - Engine for Textual Reserchers*, comunicazione presentata al *XL Congresso Internaz. di Studi della SLI: Linguistica e modelli tecnologici di ricerca, Vercelli, 21-23 settembre 2006*, ed in corso di stampa negli *Atti*.
i.p. Adriano Allora, *Per una tipologia della comunicazione mediata dalla rete. Variazione diamesica generale*, in corso di stampa in “Bollettino dell’Atlante Linguistico Italiano”.

ALLORA - BARBERA
¶ 5 Adriano Allora - Manuel Barbera, *Il problema legale dei corpora. Prime approssimazioni*, in questo volume, pp. 109-118.

ALLORA - MARELLO
i.p. Adriano Allora - Carla Marello, *“Ricarica clima”. Accorciamenti nella lingua dei newsgroup*, contributo per il *IX congresso internazionale della Società di linguistica e filologia italiana (SILFI). Prospettive nello studio del lessico italiano, Firenze 14-17 giugno 2006*, in corso di stampa negli *Atti*.

ARMSTRONG
1994 *Using Large Corpora*, edited by Susan Armstrong, Cambridge (Mass.) - London (En.), The MIT Pres, 1994 “A Bradford Book”, “ACL-MIT Press Series in Computational Linguistics” [= “Computational Linguistics” XIX (1993)$^{1-2}$].

BARBERA
2000-... Manuel Barbera, *Corpus-based Computational Linguistics Page: A Web Reference Resources Guide* (in breve: “CLR Guide”), prima ospitato a Trieste (SSLMIT) dal 28.viij.2000, poi a Stuttgart (Euralex) dal 30.j.2001, ed infine dal 28.viij.2001 su `http://www.bmanuel.org`; attualmente in aggiornamento.
2001 Manuel Barbera, *From EAGLES to CT Tagging: a Case for Re-usability of Resources*, in RAYSON et alii 2001, pp. 40-44.
2004 *in.* Manuel Barbera, *Il progetto FIRB. Stato dei lavori*, documento interno inedito, Ver. 7 aggiornata al febbraio 2004.
2004 Manuel Barbera, Schede *bmanuel.org*, *Gruppo di lavoro COFIN italiano antico, Gruppo di lavoro FIRB* e *Gruppo di lavoro Teleinsegnamento* in *Forum* TAL. *Libro bianco sul trattamento automatico della lingua* a cura di Andrea di Carlo ed Andrea Paoloni, Roma, Fondazione Ugo Bordoni, 2004, pp. risp. 126, 208, 209-210 e 211.
2007 *i.s.* Manuel Barbera, *I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo*, in “Cuadernos de filología italiana” XIV (2007) 11-32 in corso di stampa.
¶ 8 Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in questo volume, pp. 135-168.
¶ 23 Manuel Barbera, *Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra* guidelines *e prolegomeni*, in questo volume, pp. 373-388.

BARBERA - CORINO - ONESTI
¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

BARBERA - MARELLO
2000/2003 Manuel Barbera - Carla Marello, *Corpus Taurinense: italiano antico annotato in modo nuovo*, in MARASCHIO - POGGI SALANI 2003, pp. 685-693.
2003 *i.s.* Manuel Barbera - Carla Marello, *Corpo a corpo con l'inglese della* corpus linguistics*, anzi, della linguistica dei corpora*, in *Atti del Convegno Internazionale Lingua italiana e scienze, Firenze, Accademia della Crusca 6-8 febbraio 2003*, in corso di stampa.
2004 Manuel Barbera - Carla Marello, *VALICO (Varietà di Apprendimento della Lingua Italiana Corpus Online): una presentazione*, in "Didattica e linguistica dell'italiano come lingua straniera" II (2004)[4] 7-18.

BARONI et alii
2004 Marco Baroni - Silvia Bernardini - Federica Comastri - Lorenzo Piccioni - Alessandra Volpi - Guy Aston - Marco Mazzoleni, *Introducing the* La Repubblica *Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*, in AA. VV. 2004, pp. 1771-1774, disponibile online alla pagina `http://www.form.unitn.it/~baroni/publications/lrec2004/rep_lrec_2004.pdf`.

BOSC - MARELLO - MOSCA
2006 *Saperi per insegnare. Formare insegnanti di italiano per stranieri. Un'esperienza di collaborazione tra università e scuola*, a cura di Franca Bosc - Carla Marello - Silvana Mosca, Torino, Loescher 2006 "Università degli studi di Torino - Ufficio scolastico regionale per il Piemonte".

BRINO
2006 Giovanna Brino, *Problemi morfologici nell'etichettatura morfosintattica dello spagnolo. Strategie e procedure,* Università di Torino, Facoltà di Lingue, Tesi di Laurea 2004-2005.

CASAVECCHIA
2005 Sara Casavecchia, *Progettazione ed implementazione di corpora di lingua inglese basati sui newsgroup,* Università di Torino, Facoltà di Lingue, Tesi di Laurea 2004-2005.

CHRIST et alii
1999 Oliver Christ - Bruno M[aximilian] Schulze - Anja Hofmann - Esther König, *The IMS Corpus Workbench: Corpus Query Processor (CQP). User's Manual*, Stuttgart, Institut für maschinelle Sprachverarbeitung, August 16, 1999 (CQP V2.2), documento disponibile online come file HTML (`http://www.ims. uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/`), PS (`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/PS/cqpman.ps.gz`) o PDF (`http://www.ims.uni-stuttgart.de/projekte/CorpusWorknbench/CQPUserManual/PDF/cqpman.pdf`).

CHRIST - SCHULZE
1996 Oliver Christ - Bruno Maximilian Schulze, *CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora*, in FELDWEG - HINRICHS 1996; disponibile online alla pagina `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz`.

CIGNETTI
¶ 11 Luca Cignetti, *Alcune forme di polifonia testuale nei notiziari accademici di Athenaeum. Aspetti funzionali ed argomentativi*, in questo volume, pp. 199-207.

CIURCINA - RICOLFI

¶ 7 Marco Ciurcina - Marco Ricolfi, *Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora*, in questo volume, pp. 127-132.

COLOMBO

2006 Simona Colombo, *MorFO. Software di visualizzazione di morfemi in testi*, in CORINO - MARELLO - ONESTI 2006, vol. I, pp. 147-150.

CORINO

2006 Elisa Corino, *MorFo: morfemi fondamentali per capire l'italiano*, in BOSC - MARELLO - MOSCA 2006, pp. 285-296.

¶ 13 Elisa Corino, *NUNC (Newsgroup UseNet Corpora). Questioni metodologiche ed aspetti della testualità*, in questo volume, pp. 225-252.

CORINO - COLOMBO

2004 Elisa Corino - Simona Colombo, *MorFo (Morfemi Fondamentali On-line): per (far) imparare le parole italiane*, in VALENTINI et alii 2004, pp. 149-169.

CORINO - HEID - SCHAUPP

2007 *i.p.* Elisa Corino - Ulrich Heid - Annette Schaupp, *Metadaten fuer Lernerkorpora: Typen - Architektur - Abfragemoeglichkeiten, am Beispiel des VALICO-Korpus*, comunicazione al *22. DGFF-Kongress, 3-6 Oktober 2007, Sprachen lernen - Menschen bilden, Justus-Liebig-Universität Giessen*, in preparazione.

CORINO - MARELLO

2007 *i.s.* *Italiano di apprendenti. I Corpora VALICO e VINCA*, a cura di Carla Marello ed Elisa Corino, Perugia, Guerra, 2007, in corso di stampa.

CORINO - MARELLO - ONESTI

2006 *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 | Proceedings of the XII EURALEX International Congress. Torino, Italia, 6th-9th September 2006*, a cura di Elisa Corino, Carla Marello e Cristina Onesti, 2 voll., Alessandria, Edizioni dell'Orso, 2006.

FELDWEG - HINRICHS

1996 *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, herausgegeben von Helmut Feldweg und Erhard W. Hinrichs, Tübingen, Max Niemeyer Verlag, 1996 "Lexicographica. Series maior" 73.

FELDWEG - KIBIGER - THIELEN

1995 Helmut Feldweg - Ralf Kibiger - Christine Thielen, *Zum Sprachgebrauch in deutschen Newsgruppen*, in "Osnabrücker Beiträge zur Sprachtheorie" L (1995) 143-154, disponibile anche online `http://www.sfs.uni-tuebingen.de/Elwis/news.ps`.

FERRARI - MANDELLI

¶ 10 Angela Ferrari - Magda Mandelli, *Note sull'impiego dei connettivi nei notiziari accademici del corpus Athenaeum. Aspetti quantitativi e qualitativi*, in questo volume, pp. 183-198.

FITSCHEN - HEID - SCHMID

2004 Arne Fitschen - Ulrich Heid - Helmut Schmid, *SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection*, in AA. VV. 2004, pp. 1263-1266, disponibile online alla pagina `http://www.ims.uni-stuttgart.de/~schmid/`

HEID
¶ 4 Ulrich Heid, *Il* Corpus WorkBench *come strumento per la linguistica dei corpora. Principi ed applicazioni*, in questo volume, pp. 89-108.

HINRICHS et alii
1995 *Abschlußbericht [zu ELWIS Projekte]*, Projektleiter Prof Dr Erhard W. Hinrichs, Mitarbeiter Helmut Feldweg, Marie Boyle-Hinrichs und Ralf Hauser, PS file online `http://www.sfs.uni-tuebingen.de/Elwis/abschlussbericht.ps`.

KIESLER
2006 Reinhard Kiesler, *Einführung in die Problematik des Vulgärlateins*, Tübingen, Niemeyer, 2006.

KORZEN - LUNDQUIST
2007 *Comparing Anaphors between Sentences, Texts and Languages. Proceedings of the international symposium held at the Copenhagen Business School, September 1st-3rd 2005*, edited by Iørn Korzen and Lita Lundquist, Frederiksberg, Samfundslitteratur Press, 2007 "Copenhagen Studies in Language" 34.

KORZEN - MARELLO
2000 *Argomenti per una linguistica della traduzione | Notes pour une linguistique de la traduction | On Linguistics Aspects of Translation*, a cura di Iørn Korzen e Carla Marello, Alessandria, Edizioni dell'Orso, 2000 "Gli argomenti umani" 4.

LÓPEZ DÍAZ - MONTES LÓPEZ
2006 *Perspectives fonctionnelles: emprunts, économie et variations dans les langues. S.I.L.F. 2004. XXVIII Colloque de la Société internationale de linguistique fonctionnelle, tenu à Saint-Jacque-de-Compostelle et à Lugo du 20 au 26 septembre 2004*, édité par Motserrat López Díaz et María Montes López , Lugo, Editorial Axac, 2006.

MARASCHIO - POGGI SALANI
2003 *Italia linguistica anno Mille - Italia linguistica anno Duemila. Atti del XXIV Congresso internazionale di studi della Società di linguistica italiana (SLI), Firenze 19-21 ottobre 2000*, a cura di Nicoletta Maraschio e Teresa Poggi Salani, Roma Bulzoni, 2003.

MARCUS - SANTORINI - MARCINKIEVICZ
1994 Mitchell P. Marcus - Beatrice Santorini - Mary Ann Marcinkievicz, *Building a Large Annotated Corpus of English: The Penn Treebank*, in ARMSTRONG 1994, pp. 273-290. Disponibile online dalla homepage del PennTreebank al link `ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz`.

MARELLO
2007 Carla Marello, *Does Newsgroups "Quoting" Kill or Enhance Other Types of Anaphors?*, in KORZEN - LUNDQUIST 2007, pp. 145-157.

MORTARA GARAVELLI
2001 Bice Mortara Garavelli, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, Torino, Giulio Einaudi Editore, 2001 "Piccola biblioteca Einaudi. Nuova serie. Saggistica letteraria e linguistica" 100.

ONESTI
i.s. Cristina Onesti, *I corpora nella didattica delle lingue straniere*, in *Comenius 2.1 - ECNTL: European Curricula in New Technologies and Language Learning*, Amsterdam, Instituut voor de Lerarenopleiding Universiteit van Amsterdam, in stampa.

i.p. Cristina Onesti, *Identifizierung der Satzgrenzen in der Newsgroupssprache: computer- und textlinguistischen Probleme*, comunicazione alla *"Dritte Tagung Deutsche Sprachwissenschaft in Italien", Arbeitsgruppe: "Korpora und Grammatik nichtstandardisierter Sprache", Roma, 14-16 febbraio 2008*.

RAYSON et alii
2001 *Proceedings of the Corpus Linguistics 2001 Conference. Lancaster University 29 March - 2 April 2001*, edited by Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja, Lancaster, University Center for Computer Corpus Research on Language, 2001 "UCREL Technical Paper" 13.

RENZI
1998 *ITALANT: per una Grammatica dell'Italiano Antico*, a cura di Lorenzo Renzi, Padova, Centro Stampa di Palazzo Maldura, 1998.

RICOLFI
¶ iij Marco Ricolfi, *Il terribile diritto. La proprietà intellettuale: un incentivo od un ostacolo all'innovazione ed alla creatività?*, in questo volume, pp. xj-xiij.

SANTORINI
1990/1 Beatrice Santorini, *Part-of-speech Tagging Guidelines for the Penn Treebank Project*, Technical report MS-CIS-90-47, University of Pennsylvania - Department of Computer and Information Science, 1990. *3rd Revision, 2nd Printing, June 1990* è disponibile online dalla homepage del PennTreebank `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`; la *Rev. 1991 March 15* è disponibile dalla homepage del TreeTagger al link `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps`.

SCHAUPP
2006 Annette Schaupp, *Entwicklung und Anwendung eines Metadatenmodells für das italienische Lernerkorpus Valico mit Fokussierung auf den Lernerhintergrund*, Stuttgart, Institut für maschinelle Sprachverarbeitung, 2006; Diplomarbeit Nr. 51, Prüfer HD Dr. Ulrich Heid, Zweitprüfer Dr. Helmut Schmid.

SCHILLER et alii
1999 Anne Schiller - Simone Teufel - Christine Stöckert - Christine Thielen, *Guidelines für das Tagging Deutscher Textkorpora mit STTS. (Kleines und großes Tagset)*, Technical report, IMS and SfS, disponibile online alla pagina `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.ps.gz`.

SCHMID
1994 Helmut Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, paper presented at the *International Conference on New Methods in Language Processing*, Manchester (UK), 1994; versione revisionata PS/PDF online sul sito dell'IMS Stuttgart: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

SPITZER
1922/2007 Leo Spitzer, *Italienische Umgangssprache*, Bonn, Kurt Schroeder, 1922. Versione italiana: *Lingua italiana del dialogo*, a cura di Claudia Caffi e Cesare Segre, traduzione di Livia Tonelli, Milano, il Saggiatore, 2007.

TOMATIS

2004 Marco Tomatis, *Sviluppo di un modello computazionale della morfologia dell'italiano moderno applicabile a un sistema automatico di analisi basato su tecnologia FST (Finite State Transducer)*, tesi di dottorato, Torino, Dottorato in Linguistica, Linguistica applicata, Ingegneria linguistica (XVIII ciclo).

2005 Marco Tomatis, *Computational aspects of an automatic recognizer of Italian clitics*, comunicazione presentata a *COMPLEX 2005. 8th Conference on Computational Lexicography and Text Research. Hungarian Academy of Sciences - Research Institute for Linguistics, Budapest, 1068 Benczúr u. 33, 17-18 June 2005*, e stampata in *Papers in computational Lexicography. Complex 2005*, Budapest, Linguistic Institute Hungarian Academy of Sciences, 2005, pp. 223-232.

2006 *i.s.* Marco Tomatis, *SMORFIA: un analizzatore della morfologia verbale dell'italiano moderno per gli apprendenti di lingua italiana*, comunicazione presentata al *XL Congresso Internaz. di Studi della SLI: Linguistica e modelli tecnologici di ricerca, Vercelli, 21-23 settembre 2006*, ed in corso di stampa negli *Atti*.

¶ 9 Marco Tomatis, *La disambiguazione del Corpus Taurinense. Problemi teorici e pratici*, in questo volume, pp. 169-181.

VALENTINI et alii

2004 *Insegnare ad imparare in italiano L2: le abilità di studio dalla scuola all'Università. Atti del seminario Bergamo, 14-16 giugno 2004*, a cura di Ada Valentini, Rosella Bozzone Costa e Monica Piantoni, Perugia, Guerra, 2004.

VALLE

2005 *i.s.* Luca Valle, *The Retrieval of Anglicisms in Newsgroups Usenet Corpora (NUNC)*, comunicazione a *JILC 2005. 4èmes Journées Internationales de la linguistique de corpus, Lorient 15-17 septembre 2005*, Université de Bretagne Sud, 2005, in corso di stampa.

2006 Luca Valle, *Varietà diafasiche e forestierismi nell'italiano nei gruppi di discussione in rete*, in LÓPEZ DÍAZ - MONTES LÓPEZ 2006, pp. 371-374.

ZANNI

¶ 6 Samantha Zanni, *Corpora elettronici e copyright. Lo stato legale della questione*, in questo volume, pp. 119-126.

CORPORA, STRUMENTI ED ISTITUZIONI DI RIFERIMENTO.

20 Newsgroups	`http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html` `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.data.html`.
Athenaeum Corpus	`http://www.bmanuel.org/projects/at-HOME.html`
bmanuel.org	`http://www.bmanuel.org`
Centro ReTe	`http://www.rete.unito.it/`
ClitRec	`http://www.bmanuel.org/tools/cl-clitrec.htm`
CLR Guide	`http://www.bmanuel.org/clr/index.html`
corpora.unito.it	`http://www.corpora.unito.it/`.
Corpus Taurinense	`http://www.bmanuel.org/projects/ct-HOME.html`

CT → Corpus Taurinense

CWB	http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
Dima Logic	http://www.dimalogic.com/
Google Groups	http://groups.google.it/.
e-allora.net	http://www.e-allora.net/
ELWIS	http://www.sfs.uni-tuebingen.de/Elwis/
E_NT_ER	http://www.corpora.unito.it/cgi-bin/lingue/enter/enter_index.pl?corpus=VALICO
GNU	http://www.gnu.org
ILVAT	http://www.corpora.unito.it/cgi-bin/lingue/ilvat/ilvat_index.pl
IMS Stuttgart	http://www.ims.uni-stuttgart.de
Jus Jurium	http://www.bmanuel.org/projects/ju-HOME.html
MorFo	http://www.morfoweb.it
NUNC	http://www.bmanuel.org/projects/ng-HOME.html
OVI	http://www.ovi.cnr.it/
PennTreebank	http://www.cis.upenn.edu/~treebank/
“La Repubblica” C.	http://sslmit.unibo.it/repubblica
Sesamo	http://www.sesamo.it/
SfS Tübingen	http://www.sfs.uni-tuebingen.de/
SFST	http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html
SMS	http://www.e-allora.net/SMS/ms_index.php
Tree Tagger	http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
VALICO	http://www.bmanuel.org/projects/br-HOME.html
VINCA	http://www.bmanuel.org/projects/vn-HOME.html

MANUEL BARBERA[1]
(Università di Torino)

2. Il decalogo[2] della *Corpus linguistics*. *(Tanto* Esodo *20,2-17 e* Deut. *5,6-21 erano diversi)*[3].

Alles, was tief ist, liebt die Maske; die allertiefsten Dinge haben sogar einen Hass auf Bild und Gleichniss. Sollte nicht erst der Gegensatz *die rechte Verkleidung sein, in der die Scham eines Gottes einhergienge?*
Friedrich Nietzsche, *Jenseits von Gut und Böse*, § 40[4].

"ans am ieu lo chant e·l ris"
Monge de Montaudo, *L'autrier fuy en paradis*, BdT 305.12, v. 22[5].

[1] Propriamente di autore per un testo simile non se ne dovrebbe affatto parlare: troppi vi hanno contribuito. Ma chi scrive se ne assume comunque tutte le responsabilità, esimendone altrui, che è quel che conta.
[2] L'idea di stilare un decalogo della *Corpus linguistics* non è in sé nuova: ne troviamo infatti già un esempio nei *Ten commandments for computational (and corpus) linguists* di Čermák 2002, pp. 279-281, che riportiamo qui in forma abbreviata: «(1) Garbage in, garbage out. [...] (2) The more data the better. But there is never enough data to help solve everything. [...] (3) The best information comes from direct data. [...] (4) There is no all-embracing algorithm that is universal in its field and transferable to all other languages. [...] (5) Lemmatizers have invented imaginary new words, often creating non-existent entities (forms) and suggesting false ones. [...] (6) It is not all research that glitters statistically. [...] (7) Language is both regular and irregular, not everything may be captured by algorithms automatically. [...] (8) The main goal of language is to code and decode meaning. Since meaning is not limited to words only, it is wrong to concentrate on words only. [...] (9) There are no aligners that will do the job for you automatically. 99% of this has to be done manually anyway. [...] (10) It is high time to ask computational linguists what their theories and programmes cannot do, how much of the field goes by the board and is never mentioned. Their alleged comprehensive coverage may be deceptive». Tutt'ora sacrosanto; ma, a differenza dell'illustre precedente, i nostri dieci comandamenti non si propongono tanto di dare delle concrete linee guida per la costruzione di un corpus, quanto piuttosto di fare irriverentemente (al modo del Monaco medievale in epigrafe) riflettere sull'uso ed abuso della magica espressione "corpus", servendo così da introduzione paradossale al contributo sulla natura di un corpus (Barbera - Corino - Onesti, ¶ 3, in questo vol.).
[3] Lo *humour* è sana prassi consolidata della ricerca scientifica nelle culture di matrice anglosassone (per la linguistica si veda ad esempio il sagace e dissacrante *Telling the Frog Story in Academia* di Berman - Slobin 1994, p. 643): assai meno in quelle di matrice latina e segnatamente nella italiana. Inutile sperare in una inversione di tendenza se non si ha poi il coraggio di scagliare la prima pietra.
[4] Questa epigrafe è una deliberata anticipazione di quella di Hofmannstahl anteposta da Amedeo Conte al saggio che chiuderà il presente volume (Conte ¶ 22., *infra*): come per aprire e poi chiudere un cerchio.
[5] L'accostamento al corrusco apoftegma nietzschiano della solare assicurazione che Nostro Signore faceva nell'estate del 1194 (secondo credo altrove aver dimostrato) al Monge non è poi così peregrino come di primo acchito potrebbe sembrare, visto che Nietzsche stesso ebbe più volte modo di richiamarsi proprio al *gai saber* dell'Occitania medievale (cfr. ad es. *Jenseits von Gut und Böse*, § 260: «[...] den provençalischen Ritter-Dichtern [...], jenen prachtvollen erfinderischen Menschen des "gai saber", denen Europa so Vieles und beinahe sich selbst verdankt»).

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 21-23.

0 Io sono il Corpus Dio tuo
[vabbè, che vi aspettavate?]
[e non mettetevi a fondere i vitelli d'oro della Competenza, ché sennò...]
[e se proprio volete idolatrarne di più, il plurale fa "corpora"…]

1. Non avrai altro Corpus al di fuori di Me
[nessuno conosce ed usa il proprio corpus meglio del suo fattore]
[anzi, spesso solo lui riesce ad usarlo ...]
[tanto... forse che gli altri ti daranno mai il loro?]

2. Non nominare mai un Corpus invano
[occhio a quello che puoi (e devi) citare ed a quello che non puoi!]
[non fingere che sia *corpus driven* quello che vorresti che lo fosse.]
[e quattro testi che interroghi con la ricerca di Word non sono un Corpus, sono quattro testi.]

3. Ricordati di santificare le feste
[sanctae Susanna et Christina orate pro nobis.]
[non markuppate la domenica, ed il venerdì fate solo corpora di magro!]

4. Onora il padre e la madre di un Corpus
[con il sangue, la fatica e gli stenti che gli è costato, è il minimo!]
[ed un pacco di pannolini, per quanto cari, forse non è un omaggio sufficiente.]

5. Non uccidere
[per avere un Corpus ci sono mezzi ugualmente illegali, ma meno perseguibili.]
[e per avere un testo in copyright... beh, ci si può pensare.]

6. Non commettere Corpora impuri
[bilancia bene, e markuppa meglio!]
[corpora sporchi? si fanno, si fanno, ma non lo si dice troppo in giro.]

7. Non rubare testi
[che se poi ti beccano ...]
[i ricettatori di testi rischiano molto e guadagnano poco: meglio le opere d'arte.]

8. Non dire falsa testimonianza
[prima o poi se ne accorgono ...]
[i corpora c'è chi li vuole autentici: adulterati e rifatti non sono la stessa cosa.]

9. Non desiderare il Corpus altrui
[giù le mani! tanto non te lo do.]
[e poi il mio Corpus è sempre il più sexy. Oh.]
[e comunque anche gli altri non te lo danno, se non a pagamento – e che prezzi!]

10. Non desiderare la roba d'altri
[che poi di solito l'hanno rubata anche loro – e si scopiazzano pure a vicenda.]
[i tag, le DTD ed i software altrui nel tuo computer faranno esplodere il processore.]

BIBLIOGRAFIA.

BARBERA
1990 Manuel Barbera, *Saggio di edizione critica delle poesie del Monge de Montaudon*, Università degli studi di Pavia, Facoltà di Lettere e filosofia, Tesi di Laurea 1988-1989.

BARBERA - CORINO - ONESTI
¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

BERMAN - SLOBIN
1994 Ruth A. Berman - Dan Isaac Slobin, *Relating Events in Narrative: a Crosslinguistical Developmental Study*, Hillsdale (New Jersey) - Hove (UK), Lawrence Erlbaum Associates, 1994.

ČERMÁK
2002 František Čermák, *Today's Corpus Linguistics. Some Open Questions*, in "International Journal of Corpus Linguistics" VII (2002) 265-282.

CONTE
¶ 22 Amedeo Conte, *Valori normativi di verbi deontici in testi normativi*, in questo volume, pp. 363-370.

NIETZSCHE
1886/1968 Friedrich Nietzsche, *Jenseits von Gut und Böse. Vorspiel einer Philosophie der Zukunf*, 1986, in *Nietzsche. Werke. Kritische Gesamtausgabe*, begrundet von Giorgio Colli und Mazzino Montinari, weitergeführt von Volker Gerhardt, Norbert Miller, Wolfgang Müller-Lauter und Karl Pestalozzi, in Gemeinschaft mit der Berlin-Brandenburgischen Akademie der Wissenschaften, Abteilung **6** Band **2**, *Jenseits von Gut und Böse. Zur Genealogie der Moral (1886 - 1887)*, Berlin, Walter de Gruyter, 1968.

MANUEL BARBERA - ELISA CORINO - CRISTINA ONESTI[1]
(Università di Torino)

3. Cosa è un corpus?
Per una definizione più rigorosa di corpus, token, markup.

> *It is often difficult to distinguish the defining characteristics from the acquired characteristics of a form, partly because as times goes on the latter tend to became the former.*
>
> Charles Rosen, *The Classical Style: Haydn, Mozart, Beethoven*, I.2.

> *(Erinnere dich, daß wir manchmal Erklärungen fordern nicht ihres Inhalts wegen, sondern der Form der Erklärung wegen. Unsere Forderung ist eine architektonische; die Erklärung eine Art Scheingesims, das nichts trägt.)*
>
> Ludwig Wittgenstein, *Philosophische Untersuchungen*, I.217.

0. PREMESSA**.** L'idea di fondo che guida questo contributo parte abbastanza da lontano: era già stata abbozzata in una conferenza tenuta a Trieste nel maggio 2000 (Barbera 2000), ed è poi andata rafforzandosi e corroborandosi negli anni seguenti. L'aspetto più propriamente terminologico della linguistica dei corpora, dal punto di vista dell'impatto nella storia (anche futura) della lingua italiana, è stato oggetto di Barbera - Marello 2003; qui ci occuperemo invece dell'aspetto definitorio e tecnico del concetto di "corpus" e dei concetti connessi.

In genere, nell'uso corrente, al di là della nozione tradizionale di «Raccolta ordinata e completa di opere o di autori» (*DOLI* s.v., p. 706a)[2], cui non siamo qui interessati, sono spesso diffuse in italiano definizioni specialistiche assai generiche, e comunque non tecniche, che vanno[3] dalla scarna «Campione prelevato a fini scientifici dal linguista» (*DOLI* s.v., p. 706a) alla più ricca «*ling.* Raccolta di brani, singoli enunciati o altri dati linguistici, che vengono analizzati per definire la struttura di un sistema linguistico» (*DISC* s.v., p. 617b), raramente pervenendo a formulazioni più specifiche come «Raccolta di testi autentici e ricorrenti nell'uso in formato elettronico, selezionati come rappresentativi (per es.) dell'italiano corrente» (quella da cui prende le mosse Zanni ¶ 6, qui oltre).

---

[1] Il presente articolo deve le sue dimensioni all'essere progenie di quattro lavori originariamente autonomi, fusi in uno per evitare ripetizioni, e guadagnare in spazio e perspicuità. I §§ 2.0-4 sono da attribuire ad Elisa Corino, i §§ 2.5-9 a Cristina Onesti, il § 3.1 e sottoparagrafi a tutti e tre gli autori, ed il rimanente a Manuel Barbera.
[2] Per la (s)fortuna del lemma "corpus" in lessicografia cfr. Barbera - Marello 2003, nota 2 e *passim*, e qui § 3.2.
[3] O che possono anche mancare del tutto, come ad esempio nel *GDLI*, supplemento 2004 compreso.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 25-88.

Questa ultima definizione è già sufficiente per coprire molti degli usi del termine "corpus" nel discorso comune, ma non copre completamente l'accezione specialistica[4] cui fanno, implicitamente o meno, riferimento i linguisti computazionali: da questo punto di vista, infatti, una definizione di tale tipo è in realtà corretta ma incompleta. D'altra parte è un fatto che non esiste in realtà una definizione esplicita che sia completamente soddisfacente[5], ossia che da un lato copra almeno la maggior parte degli oggetti che il linguista di corpora chiama "corpus", e che dall'altro consenta *sempre* di stabilire non ambiguamente se un oggetto è o non è un "corpus"[6]. Non a caso, secondo riporta Tognini-Bonelli 2001, pp. 52-53, «every few months in the *Corpora* and other relevant lists there is a query about whether or not a certain collection of language, or a means of collecting it, would constitute a corpus, and there is a wide spectrum of views elicited by these stimuli». La trattazione di riferimento più ampia e diffusa è probabilmente McEnery - Wilson 2001, pp. 29-74[7]; ma una posizione di altrettanto rilievo (anticipando i risultati di cui al § 2 e sottoparagrafi) andrebbe attribuita anche a Sasaki - Witt 2004, p. 195 e Lemnitzer - Zinsmeister 2006, p. 40[8]. In generale, comunque, molte delle definizioni classiche fanno ricorso ai concetti di autenticità, rappresentatività, finitezza, o (meno spesso) forma elettronica. Ciascuno (ed altri) di questi elementi saranno analiticamente esaminati nei paragrafi seguenti (cfr. 2 e sgg.), ma quasi tutti patiscono eccezioni e singolarmente non riescono a *definire* (ossia discriminare) un oggetto "corpus" specifico: forse che abbiamo preso come caratteristiche "definitorie" delle caratteristiche più propriamente "acquisite" (nel senso proposto in epigrafe da Rosen) dai corpora nel corso della storia della ricerca?

Ma per giungere ad una definizione univoca crediamo sia necessario invocare, più che requisiti contenutistici, sempre intrinsecamente variabili[9], un'ulteriore restrizione formale oltre all'insufficiente "formato elettronico"; ed in generale la definizione a nostro avviso migliore dovrebbe essere più formale ("architettonica", per riprendere l'epigrafe wittgensteiniana) che sostanziale[10]. Per anticipare, una definizione provvisoriamente operativa, preliminare e sintetica ma relativamente completa, potrebbe essere del tipo seguente (una definizione più esauriente sarà fornita alla fine, nel § 4):

> *Raccolta di testi*[11] *in formato elettronico uniformemente trattati (ossia almeno tokenizzati ed addizionati di un markup adeguato) in modo da essere gestibili ed interrogabili informaticamente.*

In assenza delle due specifiche caratteristiche "tokenizzazione" e "markup", ed in generale di un "valore informatico aggiunto", non si può a nostro parere parlare di corpora ma solo, genericamente di "(raccolte di) testi elettronici", come quelle inserite, ad esempio, in un databa-

[4] Ad es., le collezioni di e-text disponibili sul web come quelle di Project Gutenberg, Progetto Manuzio, ecc., rientrerebbero facilmente nella definizione precedente, ma di fatto non sono dei veri corpora.

[5] Anche se più o meno tutti si riconoscono abbastanza nell'ultima definizione presentata.

[6] Quest'ultima condizione, in ispecie, come vedremo in contributi successivi di questo volume (cfr. soprattutto Allora - Barbera ¶ 5), è basilare se si vuole pervenire anche ad assegnare uno status giuridico preciso ai corpora.

[7] Ossia, appunto, il capitolo secondo, *What is a corpus and what is in it*.

[8] Potremmo chiamarle risp. la "formulazione di Lancaster" e la "formulazione di Tübingen".

[9] ed in qualche modo invocati funzionalmente alla teoria linguistica, pur talora nobilissima, che si vuole propugnare, come evidente dalla rassegna esemplare dei paragrafi seguenti.

[10] Hanno in ciò ragione Kilgarriff - Grefenstette 2003 p. 334 a dire che «[to] mix the question "What is a corpus?" with "What is a good corpus (for certain kinds of linguistic study)?" mudd*ies* the simple question "Is corpus x good for task y?" with the semantic question "Is x a corpus at all?"»; ma l'operazione che noi vogliamo compiere è esattamente l'opposta della loro (per cui «The semantic question then becomes a distraction [... and] may be set aside»): e cioè accantonare le istanze pratiche e funzionali per esaurire quelle definitorie.

[11] "Testo" è usato nella accezione semioticamente (e testologicamente) più vasta di «oggetto semiotico relazionale prevalentemente verbale» (Petőfi 2004, p. 22; e cfr. Petőfi 1988/96, p. 69).

se[12] più (come la LIZ) o meno (come Project Gutenberg) sofisticato. La precisazione è a mio parere di non poco conto perché determina proprio la specificità dell' "oggetto corpus" ed è essenziale per individuarne anche un profilo giuridico.

1. LO SPECIFICO FORMATO ELETTRONICO RICHIESTO. Le specifiche che postulavo erano soprattutto che il testo fosse uniformemente markuppato e tokenizzato[13]. Analizzeremo meglio e più approfonditamente cosa token (cfr. § 1.3) e markup (cfr. 1.4) siano. Per ora ci accontenteremo di introdurre delle prime definizioni semplicemente operative per vedere in generale che aspetto, in concreto, un corpus assuma per diventare tale.

Brunetto Latini, *Tesoretto*, vv. 113-134.	
versione non tokenizzata a stampa (*testo Contini*, *Poeti del Duecento*)	versione completamente tokenizzata (*testo CT*)
Lo Tesoro conenza. Al tempo che Fiorenza froria, e fece frutto, sì ch'ell'era del tutto la donna di Toscana (ancora che lontana ne fosse l'una parte, rimossa in altra parte, quella d'i ghibellini, per guerra d'i vicini), esso Comune saggio mi fece suo messaggio all'alto re di Spagna, ch'or è re de la Magna e la corona atende, se Dio no·llil contende: ché già sotto la luna non si truova persona che, per gentil legnaggio né per altro barnaggio, tanto degno ne fosse com' esto re Nanfosse.	Lo Tesoro conenza . A ÷l tempo che Fiorenza froria , e fece frutto , sì ch' ell' era de ÷l tutto la donna di Toscana (ancora che lontana ne fosse l' una parte , rimossa in altra parte , quella d' i ghibellini , per guerra d' i vicini) , esso Comune saggio mi fece suo messaggio a ÷ll' alto re di Spagna , ch' or è re de la Magna e la corona atende , se Dio no· lli ÷l contende : ché già sotto la luna non si truova persona che , per gentil legnaggio né per altro barnaggio , tanto degno ne fosse com' esto re Nanfosse .

Tav. 1a-b: Ortografia normale (non tokenizzata) vs testo tokenizzato (completamente)

Per tokenizzazione si intende grossomodo l'operazione di individuazione (in genere tramite un *blank* a destra ed a sinistra) dei token[14], ossia delle unità minime che il PC tratterà. Siccome queste, peraltro, non corrispondono sempre alle parole grafiche di un testo "tipograficamente

[12] Si badi che il concetto di *record* come 'unità di popolazione di un database' non è logicamente equivalente a quello di *token* (come risulterà evidente dal § 1.3), anche se funzionalmente può apparire assai simile.
[13] L'importanza della tokenizzazione è spesso sottovaluta nella teoria della linguistica computazionale: pensando «that the only interesting problems to be solved pertain to high-level semantics», per usare le efficaci parole di Fontenelle 2004, p. 469, «one tends to forget much too frequently that crucial questions about, for instance, the status of the apostrophe and the hyphen in French (breaking vs. non-breaking character) need to be adressed before one can tackle more difficult computational task». Qui se ne vuole invece rivendicare la natura essenziale ed ineliminabile, incardinandolo anzi come la caratteristica più discriminante di un corpus in quanto tale.
[14] Ma torneremo ancora a precisarne il concetto nel §.1.3.

composto", basta già questa operazione da sola a distinguere[15] i due oggetti. Il semplice esempio in Tav. 1, tratto dal Corpus Taurinense[16], può forse contribuire chiarire la nozione.

Si noti, tra l'altro, che anche se possono esservi varie gradazioni e sfumature di (più o meno) non tokenizzato e di (più o meno) tokenizzato, il discrimine tra le due categorie è non di meno netto e sempre tracciabile. Usando sempre il medesimo campione precedente si considerino infatti le due "forme" seguenti:

<table>
<tr><th>versione non tokenizzata ma più analitica
(testo OVI)</th><th>versione tokenizzata semplice
(testo CT senza grafoclitici)</th></tr>
<tr><td><pre>Lo Tesoro conenza.
Al tempo che Fiorenza
froria, e fece frutto,
sì ch' ell' era del tutto
la donna di Toscana
(ancora che lontana
ne fosse l' una parte,
rimossa in altra parte,
quella d' i ghibellini,
per guerra d' i vicini),
esso Comune saggio
mi fece suo messaggio
all' alto re di Spagna,
ch' or è re de la Magna
e la corona atende,
se Dio no· llil contende:
ché già sotto la luna
non si truova persona
che, per gentil legnaggio
né per altro barnaggio,
tanto degno ne fosse
com' esto re Nanfosse.</pre></td><td><pre>Lo Tesoro conenza .
Al tempo che Fiorenza
froria , e fece frutto ,
sì ch' ell' era del tutto
la donna di Toscana
(ancora che lontana
ne fosse l' una parte ,
rimossa in altra parte ,
quella d' i ghibellini ,
per guerra d' i vicini) ,
esso Comune saggio
mi fece suo messaggio
all' alto re di Spagna ,
ch' or è re de la Magna
e la corona atende ,
se Dio no· llil contende :
ché già sotto la luna
non si truova persona
che , per gentil legnaggio
né per altro barnaggio ,
tanto degno ne fosse
com' esto re Nanfosse .</pre></td></tr>
</table>

Tav. 2a-b: Testo non tokenizzato (ort. *quasi* normale) vs testo tokenizzato (parzialmente)

In 2a si ha un testo appena più articolato dello standard editoriale[17], mentre in 2b si ha una forma di tokenizzazione ridotta rispetto a quella messa in atto nel CT, dove non sono tokenizzati gli elementi grafoclitici[18] (soluzione peraltro la più frequente nei corpora in circolazione, come ad es. per ora nei nostri NUNC, Athenaeum, Jus Jurium, VALICO, ecc.). Orbene, il confine tra tokenizzato e non tokenizzato e non meno netto e facile da riconoscere tra 2a[19] e 2b di quanto non fosse tra 1a ed 1b: la possibilità di variazione delle due categorie, in altri termini, non concerne il discrimine tra le due categorie.

---

[15] L'operazione, si noti inoltre (soprattutto in ottica legale), è teoricamente reversibile, ma praticamente tale reversione è abbastanza onerosa (e con largo margine di fallibilità in zone idiosincratiche): uno scanning ex novo del testo a stampa potrebbe facilmente essere più economico ed affidabile.

[16] Gia usato anche in Barbera - Marello 2003.

[17] Ossia con apici sempre separati, punteggiatura in genere non separata, e separati i punti di clisia; laddove nella versione ortografica a stampa (1a) si avevano apici senza spazio sui gruppi proclitici ma con spazio sui gruppi tonici, scempiamento su assimilazione in clisia con punto in alto attaccato, e punteggiatura attaccata.

[18] Ossia con tutti gli apici separati, tutti gli interpuntemi separati, tutti i punti di clisia separati, ma tutti i clitici grafici *non* separati; laddove nella versione-CT (1b) si avevano tutti gli apici separati, tutti gli interpuntemi separati, tutti i punti di clisia separati, *e* tutti i clitici grafici separati.

[19] E non a caso la base dati testuale dell'OVI, che adotta la soluzione 2a, *non* è un corpus.

Per markup, invece, si intendono tutte le informazioni di carattere in qualche modo “sopra-segmentale” rispetto alla pura successione lineare dei caratteri del testo ed alla loro articolazione in token. Queste aggiunte possono ricoprire caratteristiche del testo, come i “registri di enfasi” (resi in tipografia con i vari corsivi, grassetti, o con i diversi tipi di carattere) e la struttura paragrafematica[20], o caratteristiche dell’edizione di quel testo, come ad esempio i numeri di pagina e di riga, o fornire informazioni esterne al testo (ad esempio dati anagrafici dell’autore).

Come, più in dettaglio, tutte queste informazioni abbastanza eterogenee e comunque extra-testuali, si possano (anche teoricamente) articolare ed organizzare, lo vedremo meglio nel § 1.3. Ora, giusto per dare un’idea concreta, il markup minimo (in formato non XML) richiesto dall’esempio del CT precedentemente utilizzato sarebbe il seguente:

versione markuppata e tokenizzata (*testo CT*)	
`@BrunettoLatini@@Tesoretto@@@Did`	`per guerra d' i vicini ) ,`
`%001`	`esso Comune saggio`
`$0175$ &V`	`mi fece suo messaggio`
`[...]`	`a ÷ll' alto re di Spagna ,`
`Lo Tesoro conenza .`	`ch' or è re de la Magna`
`A ÷l tempo che Fiorenza`	`e la corona atende ,`
`froria , e fece frutto ,`	`se Dio no· lli ÷l contende :`
`sì ch' ell' era de ÷l tutto`	`ché già sotto la luna`
`la donna di Toscana`	`non si truova persona`
`( ancora che lontana`	`che , per gentil legnaggio`
`ne fosse l' una parte ,`	`né per altro barnaggio ,`
`$0180$ rimossa in altra parte ,`	`tanto degno ne fosse`
`quella d' i ghibellini ,`	`com' esto re Nanfosse .`
markup: *@autore* *@@titolo* *@@@genere*	
%capitolo *$pagina* *&v verso*	

Tav. 3: Testo markuppato (non XML) e tokenizzato

Un tipo speciale, ma molto importante, di “informazione aggiunta” è, inoltre, quella che viene di solito chiamata *tagging*: anche se, propriamente, non è altro che un tipo particolare di markup, è usuale (ed in effetti utile) distinguerlo dal markup vero e proprio.

Il tagging consiste nell’aggiungere al testo informazioni di carattere linguistico, come le associazioni di lemma (“lemmatizzazione”), le attribuzioni di parti del discorso e categorie morfosintattiche (“POS-tagging”[21]), le segmentazioni sintattiche (con diverso grado di accuratezza, e diverse implicazioni teoriche, “chunking” e “parsing”), ecc. Nel caso dell’esempio precedente, la versione finale markuppata e taggata[22] (per lemmi, parti del discorso e categorie morfosintattiche) è diventata nel formato-CT come in Tav. 4.

[20] È tuttavia uso consolidato trattare tutti i segni di interpunzione come appartenenti a pieno titolo al testo lineare vero e proprio, identificandoli pertanto come token, e non rappresentandoli come markup.

[21] Per maggiori dettagli sul POS-tagging cfr. qui Barbera ¶¶ 8 e 23.

[22] Non forniamo qui una chiave di interpretazione dei tag numerici aggiunti al testo, bastando rimandare il lettore alla presentazione che ne è fatta oltre in questo volume (cfr. Barbera ¶ 8).

```
@BrunettoLatini@@Tesoretto@@@Did
%001 $0175$ &V
[...]Lo_lem=lo,60,0,4,6,0,0 Tesoro_lem=tesoro,20,0,4,6,0,0
    conenza_lem=cominciare,111,3,0,6,0,0 ._lem=stop,70,0,0,0,0,0
A_lem=a,56,0,0,0,0,0 ÷l_lem=il,60,0,4,6,0,0 tempo_lem=tempo,20,0,4,6,0,0
    che_lem=che,36,0,4;5,6;7,0,0 Fiorenza_lem=firenze,21,0,5,6,0,0
froria_lem=fiorire,112,3,0,6,0,0 ,_lem=comma,71,0,0,0,0,0 e_lem=e,50,0,0,0,0,0
    fece_lem=fare/-si/,113,3,0,6,0,0 frutto_lem=frutto,20,0,4,6,0,0
    ,_lem=comma,71,0,0,0,0,0
sì_lem=sì,45,0,0,0,8,0 ch'_lem=che,51,0,0,0,0,0 ell'_lem=ella,37,3,5,6,0,0
    era_lem=essere,212,3,0,6,0,0 de_lem=di,56,0,0,0,0,0 ÷l_lem=il,60,0,4,6,0,0
    tutto_lem=tutto,32,0,4,6,0,0
la_lem=la,60,0,5,6,0,0 donna_lem=donna,20,0,5,6,0,0 di_lem=di,56,0,0,0,0,0
    Toscana_lem=toscana,21,0,5,6,0,0
(_lem=parenleft,71,0,0,0,0,0 ancora_lem=ancora,45,0,0,0,8,0 che_lem=che,51,0,0,0,0,0
    lontana_lem=lontano,26,0,5,6,8,0
ne_lem=ne,31,0,0,0,0,0 fosse_lem=essere/-si/,216,3,0,6,0,0 l'_lem=la,60,0,5,6,0,0
    una_lem=uno,61,0,5,6,0,0 parte_lem=parte,20,0,5,6,0,0 ,_lem=comma,71,0,0,0,0,0
$0180$ rimossa_lem=rimuovere,123,0,5,6,0,0 in_lem=in,56,0,0,0,0,0
    altra_lem=altro,26,0,5,6,8,0 parte_lem=parte,20,0,5,6,0,0
    ,_lem=comma,71,0,0,0,0,0
quella_lem=quello,30,0,5,6,0,0 d'_lem=di,56,0,0,0,0,0 i_lem=il,60,0,4,7,0,0
    ghibellini_lem=ghibellino,20,0,4,7,0,0 ,_lem=comma,71,0,0,0,0,0
per_lem=per,56,0,0,0,0,0 guerra_lem=guerra,20,0,5,6,0,0 d'_lem=di,56,0,0,0,0,0
    i_lem=il,60,0,4,7,0,0 vicini_lem=vicino,20,0,4,7,0,0
    )_lem=parenright,71,0,0,0,0,0 ,_lem=comma,71,0,0,0,0,0
esso_lem=esso,30,0,4,6,0,0 Comune_lem=comune,20,0,4,6,0,0
    saggio_lem=saggio,26,0,4,6,8,0
mi_lem=mi,39,1,4;5,6,0,0 fece_lem=fare/-si/,113,3,0,6,0,0
    suo_lem=suo,33,3,4;5,6;7,0,0 messaggio_lem=messaggio,20,0,4,6,0,0
a_lem=a,56,0,0,0,0,0 ÷ll'_lem=lo,60,0,4,6,0,0 alto_lem=alto,26,0,4,6,8,0
    re_lem=re,20,0,4,6,0,0 di_lem=di,56,0,0,0,0,0 Spagna_lem=spagna,21,0,5,6,0,0
    ,_lem=comma,71,0,0,0,0,0
ch'_lem=che,36,0,4;5,6;7,0,0 or_lem=ora,45,0,0,0,8,0 è_lem=essere/-si/,211,3,0,6,0,0
    re_lem=re,20,0,4,6,0,0 de_lem=di,56,0,0,0,0,0 la_lem=la,60,0,5,6,0,0
    Magna_lem=magna,21,0,5,6,0,0
e_lem=e,50,0,0,0,0,0 la_lem=la,60,0,5,6,0,0 corona_lem=corona,20,0,5,6,0,0
    atende_lem=attendere,111,3,0,6,0,0 ,_lem=comma,71,0,0,0,0,0
se_lem=se,51,0,0,0,0,0 Dio_lem=dio,21,0,4,6,0,0 no·_lem=non,45,0,0,0,8,0
    lli_lem=lo,60,0,4,7,0,0 ÷l_lem=lo,39,3,4,6,0,0
    contende_lem=contendere,111,3,0,6,0,0 :_lem=colon,71,0,0,0,0,0
ché_lem=ché,51,0,0,0,0,0 già_lem=già,45,0,0,0,8,0 sotto_lem=sotto,56,0,0,0,0,0
    la_lem=la,60,0,5,6,0,0 luna_lem=luna,20,0,5,6,0,0
non_lem=non,45,0,0,0,8,0 si_lem=si,39,3,4;5,6;7,0,0 truova_lem=trovare/-
    si/,111,3,0,6,0,0 persona_lem=persona,20,0,5,6,0,0
che_lem=che,36,0,4;5,6;7,0,0 ,_lem=comma,71,0,0,0,0,0 per_lem=per,56,0,0,0,0,0
    gentil_lem=gentile,26,0,4,6,8,0 legnaggio_lem=lignaggio,20,0,4,6,0,0
né_lem=né,50,0,0,0,0,0 per_lem=per,56,0,0,0,0,0 altro_lem=altro,26,0,4,6,8,0
    barnaggio_lem=barnaggio,20,0,4,6,0,0 ,_lem=comma,71,0,0,0,0,0
tanto_lem=tanto,45,0,0,0,8,0 degno_lem=degno,26,0,4,6,8,0 ne_lem=ne,31,0,0,0,0,0
    fosse_lem=essere/-si/,216,3,0,6,0,0
com'_lem=come,51,0,0,0,0,0 esto_lem=esto,30,0,4,6,0,0 re_lem=re,20,0,4,6,0,0
    Nanfosse_lem=nanfosse,21,0,4,6,0,0 ._lem=stop,70,0,0,0,0,0
```

Tav. 4: Testo markuppato, tokenizzato, lemmatizzato e POS-taggato (formato-CT)

È importante sottolineare che solo di fronte a testi preparati con tokenizzazione e markup elementare (non necessariamente anche con tagging[23]) si può parlare di *corpus*: ossia, dei testi

[23] La maggior parte dei corpora in circolazione, in effetti, sono non taggati; ed anzi è proprio ai *raw corpora*, completamente non taggati, che si affida l'ala più puristica ed estremista della disciplina: «in corpus-driven linguistics you do not use pre-tagged text, but you process the raw text directly and then the patterns of this uncontaminated text are able to be observed» (Sinclair 2000, p. 36). La più ampia e circostanziata discussione dei due procedimenti, *corpu -based* (pp. 64-83) e *corpus-driven* (84-100), è Tognini-Bonelli 2001, pp. 64-100.

in cui fossero state implementati solo tokenizzazione ed il markup ma non il tagging, sarebbero già un corpus; mentre collezioni di testi (come l'OVI[24], per appunto ad esempio), pur "machine-readable", interrogabili, e, comunque, ugualmente appartenenti alla categoria legale delle "banche dati", ma in cui manca anche solo la tokenizzazione, non sono dei corpora[25].

1.1 LA NATURA "IBRIDA" DEL CORPUS. Possiamo ora comprendere meglio la particolare natura "ibrida" di un corpus.

Da un lato, è vero, un corpus richiede l'immissione di testi al pari di una qualsiasi opera collettiva (antologie, ecc.), dall'altro lato, però, richiede anche l'immissione di tutta una serie di procedure e marcature che sono informatiche non meno dei listati di cui si compone un programma. L'oggetto finale non è più equiparabile ad un mero testo (di fatto il testo di partenza non esiste più nella sua forma primitiva) ma neanche *tout court* ad un software convenzionale.

Questo è ancora più evidente se si vede la forma che il corpus così preparato assume nel formato CQP[26], dove ad ogni token (e ad ogni tag *XML-like* del markup[27]) è assegnata una riga distinta, e ad ogni fascia di annotazione una colonna, come si vede dalla tavola 5 *infra*, che riporta l'inizio del medesimo lacerto di cui alle tavole 1-4 nel formato CQP (quasi) finale[28].

La definizione legale stessa di banca dati, che pure è l'unica applicabile ai corpora (cfr. *infra* Zanni ¶ 6), ossia

> *«raccolte di opere, dati o altri elementi indipendenti sistematicamente o metodicamente disposti ed individualmente accessibili mediante mezzi elettronici o in altro modo»*
>
> (dlgs n. 169 del 1999, art. 2 comma 1[29]),

non è pertanto neanch'essa completamente adeguata all'oggetto perché non ne coglie la natura intrinsecamente informatica (e non solo l' "accessibilità" informatica). E di fatto non è una definizione individuante specificamente l'oggetto corpus, costituendone semmai, come abbiamo visto, un iperonimo.

---

[24] Nel suo sito l'OVI si riferisce talvolta per brevità a quello che correttamente definisce "database testuale dell'Opera del Vocabolario Italiano" come "corpus testuale dell'OVI": l'uso di "corpus" è chiaramente informale.

[25] Il riferimento alla nozione di "rappresentatività", a volte invocato, non pare da solo sufficiente (cfr. § 2.3).

[26] Per di più, si badi, questo è il formato visualizzabile del CQP, non quello che materialmente è usato dal software, che è esclusivamente indicizzato: puramente informatico, quindi.

[27] Per contenere le dimensioni della tavola alcuni tag sono tuttavia stati raccolti sulla stessa riga, segnalando l'accapo con il segno <|>.

[28] Non sono infatti ancora state inserite le fasce filologiche e delle multiword, per cui cfr. oltre tav. 7.

[29] Il decreto modificativo della legge 22 aprile 1941, n.633, che recepisce (con minime varianti formali qui evidenziate dal corsivo) il testo della direttiva europea: «raccolt*a* di opere, dati o altri elementi indipendenti sistematicamente o metodicamente disposti ed individualmente accessibili *grazie a* mezzi elettronici o in altro modo» (Dir. CE 96/9 art. 1 comma 2).

```
<author BrunettoLatini> | <title Tesoretto>  | <genr Did> | <chapter 001> |
<page 0175> | <type verse>  | [...] | <s 1429> | <line 263>
Lo                lo            |art.d|           |60,0,4,6,0,0|       V   Did
Tesoro            tesoro        |n.c|             |20,0,4,6,0,0|       V   Did
conenza           cominciare    |v.m.f.ind.pr|    |111,3,0,6,0,0|      V   Did
.                 stop          |punct.fi|        |70,0,0,0,0,0|       V   Did
</s> | <s 1430> | </line> | <line 264>
A                 a             |adp.pre|         |56,0,0,0,0,0|       V   Did
÷l                il            |art.d|           |60,0,4,6,0,0|       V   Did
tempo             tempo         |n.c|             |20,0,4,6,0,0|       V   Did
che               che           |pd.rel|          |36,0,4;5,6;7,0,0|   V   Did
Fiorenza          firenze       |n.p|             |21,0,5,6,0,0|       V   Did
</line> | <line 265>
froria            fiorire       |v.m.f.ind.ipf|   |112,3,0,6,0,0|      V   Did
,                 comma         |punct.nfi|       |71,0,0,0,0,0|       V   Did
e                 e             |conj.co|         |50,0,0,0,0,0|       V   Did
fece              fare/-si/     |v.m.f.ind.pt|    |113,3,0,6,0,0|      V   Did
frutto            frutto        |n.c|             |20,0,4,6,0,0|       V   Did
,                 comma         |punct.nfi|       |71,0,0,0,0,0|       V   Did
</line> | <line 266>
sì                sì            |adv.gn|          |45,0,0,0,8,0|       V   Did
ch'               che           |conj.sb|         |51,0,0,0,0,0|       V   Did
ell'              ella          |pd.per.s.no|     |37,3,5,6,0,0|       V   Did
era               essere        |v.a.f.ind.ipf|   |212,3,0,6,0,0|      V   Did
de                di            |adp.pre|         |56,0,0,0,0,0|       V   Did
÷l                il            |art.d|           |60,0,4,6,0,0|       V   Did
tutto             tutto         |pd.ind|          |32,0,4,6,0,0|       V   Did
</line> | <line 267>
la                la            |art.d|           |60,0,5,6,0,0|       V   Did
donna             donna         |n.c|             |20,0,5,6,0,0|       V   Did
di                di            |adp.pre|         |56,0,0,0,0,0|       V   Did
Toscana           toscana       |n.p|             |21,0,5,6,0,0|       V   Did
</line> | <line 268>
(                 parenleft     |punct.nfi|       |71,0,0,0,0,0|       V   Did
ancora            ancora        |adv.gn|          |45,0,0,0,8,0|       V   Did
che               che           |conj.sb|         |51,0,0,0,0,0|       V   Did
lontana           lontano       |adj|             |26,0,5,6,8,0|       V   Did
</line> | <line 269>
ne                ne            |adv.pc|          |46,0,0,0,0,0|       V   Did
fosse             essere/-si/   |v.a.f.sub.ipf|   |216,3,0,6,0,0|      V   Did
l'                lo            |art.d|           |60,0,4,6;7,0,0|     V   Did
una               uno           |pd.ind|          |32,0,5,6,0,0|       V   Did
parte             parte         |n.c|             |20,0,5,6,0,0|       V   Did
,                 comma         |punct.nfi|       |71,0,0,0,0,0|       V   Did
</page> | <page 0180>
rimossa           rimuovere     |v.m.nf.par.pt|   |123,0,5,6,0,0|      V   Did
in                in            |adp.pre|         |56,0,0,0,0,0|       V   Did
altra             altro         |pd.ind|          |32,0,5,6,0,0|       V   Did
parte             parte         |n.c|             |20,0,5,6,0,0|       V   Did
,                 comma         |punct.nfi|       |71,0,0,0,0,0|       V   Did
</line> | <line 270>
quella            quello        |pd.dem.s|        |30,0,5,6,0,0|       V   Did
d'                di            |adp.pre|         |56,0,0,0,0,0|       V   Did
i                 il            |art.d|           |60,0,4,7,0,0|       V   Did
ghibellini        ghibellino    |n.c|             |20,0,4,7,0,0|       V   Did
,                 comma         |punct.nfi|       |71,0,0,0,0,0|       V   Did
</line> | <line 271>
```

Tav. 5: Testo markuppato, tokenizzato, lemmatizzato e POS-taggato (formato-CQP)

1.2 I CORPORA PREISTORICI. Un effetto di questa impostazione è che si possono tecnicamente considerare corpora solo "oggetti" nati dagli anni Sessanta in poi: nell'era, cioè, dei computer, che, per la nostra disciplina, potremmo ben chiamare post-Brown[30]. Infatti storicamente

> *«corpus linguistics today is so thoroughly dependent on computers that it would be easy to suppose that the discipline only began when computers had become available to linguists. That is by no means true. We saw [...] that some work distantly related to corpus linguistics was happening a very long time ago. But the man who really inaugurated the modern corpus-linguistics tradiction was Charles Fries, who worked in the 1950s – a time when digital computers were primitive machines familiar only to a scattering of the world's mathematicians. Fries compiled a spoken English corpus by recording about 250,000 words of telephone conversations. He used this as the basis for a published description of the English structure, which aimed to reflect the language as it actually is used spontaneously, rather than as the philological tradition says that it is or should be used.»*
>
> Sampson 2004a, p. 9.

Gli "oggetti" precedenti, come quelli approntati dal Fries[31], che potremmo ben qualificare come corpora preistorici[32] (o *precorpora*) restano in effetti al di fuori dai paletti della nostra definizione di quei corpora che pure di essi sono i naturali discendenti[33]. Questo è indubbiamente un inconveniente, ma l'aporia è forse solo apparente: è naturale, infatti, che i bonobo (*Pan paniscus*) siano nostri strettissimi parenti, ma è altrettanto naturale che vi sia tra loro e noi (*Homo sapiens*) una ferma barriera riproduttiva. Né ciò ci par strano: e perché allora dovrebbe parerci strano porre una barriera definitoria tra gli antenati dei corpora ed i corpora stessi?

Naturalmente, inoltre, ciò non inficia affatto l'importante operazione culturale attuata da Sampson, collegando l'attuale linguistica empirica a tutta la tradizione linguistica pre- e non-generativa, vuoi, popperianamente, dal punto di vista "politico" (Sampson 1979), vuoi da quello epistemologico (Sampson 1997), vuoi da quello più eminentemente fattuale (Sampson 2001): quello che qui è in gioco, infatti, non è la storia della linguistica empirica, né la sua interpretazione o valutazione, ma semplicemente (ed assai più modestamente) la definizione tecnica di quel particolare oggetto che è un corpus.

Anche senza tutte le ricche implicazioni della studiata mossa sampsoniana, va comunque segnalato che nel tracciare la storia della linguistica dei corpora «c'è la diffusa tendenza a parlare di studi di *corpus linguistics* per tutti gli studi basati su dati empirici anche prima di quaranta anni fa» (Barbera - Marello 2003 *i.s.*). In effetti, quello che il moderno linguista dei corpora fa non è intrinsecamente nulla di diverso da quello che il filologo-linguista di fine Ottocento faceva sui testi, o che il linguista strutturale (saussuriano, funzionalista o bloomfieldiano che fosse) della prima metà del Novecento faceva su collezioni di *paroles*: la differenza, materialmente specifica, è solo l'uso delle schedine cartacee anziché dei computer. Il diverso valore attribuito alla statistica (che ne è la differenza teorica forse più rilevante) ne è solo un portato[34].

---

[30] Dal nome del capostipite di tutti i corpora attuali, il Brown Corpus of American Written English, compilato da Winthrop Nelson Francis ed Henry Kučera alla Brown University del Rhode Island e pubblicato nel 1964.

[31] Il riferimento è soprattutto Fries 1952, ma Fries, ricordo, era della classe 1887...

[32] Ormeggiando il titolo di un noto articolo di Francis, *Language Corpora B.C.* (Francis 1992).

[33] La pionieristica opera del padre Busa su Tommaso d'Aquino, già fondata su spogli elettronici, si porrebbe invece al limite delle due epoche, confermando l'idea (Marello 1996, pp. 167-8) che sia proprio Busa che debba essere considerato il vero capostipite della nostra *gens*. Capostipite (classe del '13) peraltro tuttora ben presente ed attivo: se l'incontro del padre con Watson all'IBM di New York nel 1949 fa ormai parte dell'epopea, così come il suo primo *Saggio* (Busa 1951), la versione online del fondamentale *Index Thomisticus* è infatti del 2005.

[34] In una corretta visione storiografica, il vero punto di rottura e novità nella tradizione della linguistica moderna è in realtà la nascita e lo sviluppo della teoria generativa: prima, intorno e dopo quella grande avventura si ha sostanziale continuità, anche se certo la sua "novità" è stata talvolta fin troppo esibita da Chomsky, mettendo in

Così «Svartvik 1992a, p. 7, risale addirittura a Otto Jespersen» (Barbera - Marello 2003 *i.s.*); ed analogamente anche Meyer 2002, p. xij. Ed, oltre a Jespersen[35] ed a Fries, è soprattutto Bloomfield[36] ad essere chiamato anche in causa: «an empiricist corpus-based approach is perhaps even more clearly seen in the works of American structuralists (the 'post-Bloomfieldians'), particularly Zellig Harris. For example, (Harris 1951) is an attempt to find discovery procedures by which a language's structure can be discovered automatically» (Manning - Schütze 1999, p. 6)[37]. Non mancano anche operazioni di più ampio respiro, come Lüdeling - Kytö 2007 *i.s.*, che si riconnette addirittura alla nascita della linguistica moderna *tout court*[38], ed alla glottologia ottocentesca e neogrammatica in particolare. Altrettanto, se non più, radicale è poi l'operazione di Sabatini 2006 che riconnette la linguistica dei corpora a quella lessicografia basata concretamente su testi come teorizzata e praticata dall'Accademia della Crusca fin dalla fine del Cinquecento (in opposizione, ad esempio, alla pratica lessicografica di Francia), od alla storia stessa (Sabatini ¶ ij) della lingua italiana, visto come una sorta di "lingua *corpus-based*".

Ma, ad ammissione dello stesso Leech (per cui cfr. n.37), «there is virtually a discontinuity between the corpus linguists of that era and the later variety of corpus linguists» (Leech 1991, p. 8): questa cesura la attribuiva ai fasti della meteora generativa, ed i fondatori della «new school of corpus linguistics» erano ravvisati, ai due lati dell'Atlantico, in Randolph Quirck, ed in Nelson Francis e Henry Kučera; ed analogamente Svartvik 1992a, p. 7: «While the manual excerpting of textual data has been the regular means of gathering information for linguistic description, its modern form, which only recently has come to be known by the name of corpus linguistics – the use of large collections of text available in machine-readable form – only dates back to the early 1960s. The beginning of it all was the making of the Brown Corpus, "a standard sample of present-day English for use with digital computers"». Opinione, per dipiù, condivisa anche da un recente studio bibliografico di riferimento come Lenz 2000, p. 6: «Die Disziplin in modernen Sinne [...] nahm ihren Anfang mit der Verfügbarkeit der ersten Korpora, einheitlich kodierter elektronisch verfügbarer Textsammlungen, in den sechziger Jahren. Als Urkorpus gilt hier das Brown Corpus [...]».

In effetti, la scelta di far partire la fase "storica" della linguistica dei corpora dalla (ri)fondazione degli anni Sessanta con Francis e Kučera, per quanto invalsa e storiograficamente corretta sia la detta tendenza nella linguistica e nella lessicografia a base empirica ad acquisirsi antenati illustri ed assicurarsi discendenza da magnanimi lombi, non solo non è affatto isolata, ma si può anzi considerare lo standard medio[39], che pur recependo quella sostanziale "continuità extra-generativa" cui accennavamo nella nota 37, percepisce nell'apparizione del Brown corpus un vero discrimine epocale.

ombra i debiti immediati con Frege e lo Husserl delle *Ricerche logiche*, da un lato, e con Jespersen, dall'altro, a favore di un non del tutto probabile "salto" all'indietro nel Settecento cartesiano.

[35] Antenato la cui grandezza (e sotterranea centralità nella storia della linguistica) è implicitamente confermata dallo strano fatto di essere considerato possibile avo tanto dalla tradizione generativa quanto da quella empirica.

[36] Un collegamento con il funzionalismo di matrice hallidayana è invece proposto da Thompson - Huston 2005.

[37] Questa linea era stata individuata soprattutto da Geoffrey Leech (sulla scorta della tesi, inedita, di Marc Sebba del 1989), che si chiedeva, appunto, se l'inizio della *corpus linguistics* andasse ricondotto «to the era of post-Bloomfieldian structural linguistics in the USA» (Leech 1991, p. 8), visto che «this was when linguists (such as Harris and Hill in the 1950s) were under the influence of a positivist and behaviourist view of the science, and regarded the 'corpus' as the primary explicandum of linguistics» (*ibidem*).

[38] «As a methodology, the rise of modern corpus linguistics is closely related to the history of linguistics as an empirical science. Many techniques that are in use in corpus linguistics are much older than electronic computers: many of them are rooted in the tradition of the late 18th and 19th century when linguistics was for the first time claimed to be a 'real', or empirical, science. Modern corpus linguistics has used and developed these methods» (Lüdeling - Kytö - McEnery [2006]).

[39] Così ad es. Johansson 1991; la più rara posizione opposta (ossia mischiare indistintamente collezioni non elettroniche di testi e corpora elettronici), può essere invece esemplificata con Kennedy 1998 (cfr. la cit. in § 2.8).

L'ultima obiezione, infine, che si potrebbe fare al requisito informatico per la definizione "stretta" di corpus (su cui cfr. la presentazione analitica nel § 2.8 *infra*), ed alla conseguente periodizzazione storica qui proposta, è che "oggetti" non informatici sono usati non solo in quella che abbiamo chiamata l' "epoca preistorica" della *corpus linguistics*, ma anche nell'epoca attuale (*adcorpora*), in quanto, almeno per gli studi storici, valgono ancora a tutti gli effetti come "corpora" i testi cartacei che non sempre vengono resi in *machine-readable form* perché già altrimenti indicizzati (Frank Henrik Müller, *c.p.*; cfr. anche Kopotev 2003, p. 35 [ver. inglese]: «old corpuses (i.e., simply electronic texts) still retain their linguistic value in many areas»). La nostra risposta è che non si vede perché un linguista dei corpora non debba occuparsi anche di altri oggetti (e servirsi di altri strumenti) oltre che dei corpora propriamente detti. D'altra parte, che, per tradizione e convenienza, ci si occupi anche di oggetti di studio che non sono propriamente quelli specifici che dànno il nome alla propria disciplina, è fenomeno abbastanza frequente nelle scienze[40], e non provoca nessuna particolare perplessità nella comunità scientifica.

1.3 LA TOKENIZZAZIONE: TOKEN E TYPE. Data l'importanza del concetto, non è forse male spendere qualche parola in più sul concetto di token, considerandone anche la storia e gli aspetti teorici; anche perché, come già si è osservato «in linguistic textbooks tokenization is quickly dispatched as a relatively uninteresting preprocessing step performed before linguistic analysis ids undertaken. In reality, tokenization is a non-trivial problem» (Grefenstette - Tapanainen 1994, p. 79).

Per token, abbiamo detto, si intendono di solito le unità minime in cui è diviso il testo elettronico (che per il computer è solo una lunga stringa di caratteri). «The isolation of word-like units from a text is called tokenization» (Grefenstette - Tapanainen 1994, p. 79): in altre parole «*token* means the individual appearance of a word in a certain position in a text. For example, one can consider the wordform *dogs* as an instance of the word *dog*. And the wordform *dogs* that appears in, say, line 13 of page 143 as a specific *token*» (Grefenstette 1999, p. 117; cfr. anche Mikheev 2003). Il type, a sua volta sarebbe in prima approssimazione il descrittore della classe di tutti i token identici[41], così come il lemma è il descrittore della classe di tutti i type appartenenti allo stesso paradigma lessicale.

La tokenizzazione (*tokenization*), propriamente, è dunque la serie di operazioni necessarie per rendere ogni "parola" od elemento significante del testo (come, in direzione intraverbale, i grafoclitici e, in direzione extraverbale, le multiword) visibile come token dalla macchina, tipicamente individuandolo con spazi prima e dopo.

Varie strategie sono state elaborate per automatizzarne il più possibile la procedura, da più sofisticati moduli direttamente inseriti nei tagger a più semplici applicazioni AWK (cfr. Brennan 2000) o LEX (cfr. Grefenstette - Tapanainen 1994 e Grefenstette 1999). Nel caso del Corpus Taurinense sopra presentato, ad esempio, data la scarsa dimensione del corpus ed il suo alto tasso di variazione ortografica[42], si era preferito procedere in modo semimanuale, affron-

---

[40] Un esempio fra molti: i funghi (cfr. Ainsworth - Bisby 1995) sono oggi intesi come composti da soli quattro phyla (*Ascomycota*, *Basidiomycota*, *Chytridiomycota* e *Zygomycota*) ma da sempre, ieri come oggi, i micologi studiano anche phyla cladisticamente assai distanti ed irrelati come gli *Oomycota* (che, anzi, comprendendo un ordine fitopatologicamente importante come le *Peronosporales*, sono assai studiati non fosse che per il loro rilievo economico) od i diversi componenti del raggruppamento polifiletico "*Myxomicota*".

[41] In realtà le cose sono un po' più complesse, come ben evidenzia Quine 1987 cit. qui avanti, e come è già peraltro inferibile dal passo, fondante, di Peirce riportato qui sotto. Per gli scopi della *corpus linguistics*, comunque, questa prima approssimazione è spesso stata giudicata sufficiente (cfr. ad esempio l'accezione con cui *token* è usato nei *Wordsmith's Tools*, uno dei software più diffusi nel settore).

[42] Non solo l'italiano antico presenta, infatti, un margine di oscillazione ortografica molto più alto di quello generalmente incontrato nella tokenizzazione dei principali corpora di lingue moderne su cui si è finora prevalentemente lavorato, ma inoltre gli editori dei singoli testi componenti il corpus hanno anche usato criteri di

tando frazionatamente ogni singolo problema, e ricorrendo a piccole *routines* GAWK solo per sostituzioni puntuali. Decidere cosa in un testo debba essere un token, ed individuarlo conformemente, non è spesso facile: ma su questo le considerazioni di Grefenstette - Tapanainen 1994 rimangono, tredici anni dopo, ancora validissime ed ormai acquisite. È forse invece della natura teorica dei concetti di type e token che non sembra esserci una diffusa consapevolezza nella comunità dei linguisti, sicché merita di spendervi qualche ulteriore parola.

La coppia type e token, infatti, ha anche una solida portata teorica in semiotica, logica e filosofia del linguaggio[43]. La loro introduzione risale infatti a Charles Sanders Peirce che nel 1906, nei *Prolegomena to an Apology for Pragmaticism*, ne dava una definizione illuminante, anche linguisticamente:

> *«A common mode of estimating the amount of matter in a MS. or printed book is to count the number of words. There will ordinarily be about twenty* the's *on a page, and of course they count as twenty words. In another sense of the word "word", however, there is but one word "the" in the English language; and it is impossible that this word should lie visibly on a page or be heard in any voice, for the reason that it is not a Single thing or Single event. It does not exist; it only determines things that do exist. Such a definitely significant Form, I propose to term a* Type. *A Single event which happens once and whose identity is limited to that one happening or a Single object or thing which is in some single place at any one instant of time, such event or thing being significant only as occurring just when and where it does, such as this or that word on a single line of a single page of a single copy of a book, I will venture to call a* Token. *An indefinite significant character* such *as a tone of voice can neither be called a Type nor a Token. I propose to call such a Sign a* Tone. *In order that a Type may be used, it has to be embodied in a Token which shall be a sign of the Type, and thereby of the object the Type signifies. I propose to call such a Token of a Type an* Instance *of the Type»*
>
> Peirce 1906/31-58, p. 537 (anche in *Commens Dictionary*, s.v.)

Una ulteriore messa a punto, esemplarmente ben chiara, è poi venuta, ancora una volta, dal fronte filosofico, dove il grande Willard van Orman Quine spiegava la questione in termini linguisticamente assai appropriati:

> *«*ES IST DER GEIST DER SICH DEN KÖRPER BAUT: *such is the nine-word inscription on a Harward* museum. *The count is nine because we count* der *both times; we are counting concrete physical objects, nine in a row. When on the other hand statistics are compiled regarding students' vocabulaires, a firm line is drawn at repetitions; no cheating. Such are two contrasting senses in which we use the word* word. *A word in the second sense is not a physical object, not a dribble of ink or an incision in granite, but an abstract object. In this second sense of the word* word *it is not two words* der *that turn up in the inscription, but one word* der *that gets inscribed twice. Words in the first sense have come to be called* tokens*; words in the second sense are called* types.*»*
>
> Quine 1987, pp. 216-7.

normalizzazione grafica spesso molto diversi tra loro; né la modesta consistenza del corpus avrebbe d'altro canto consentito di ammortizzare tali inconvenienti con i grandi numeri

[43] Circostanza, tra l'altro, che ne sconsiglia la rinuncia terminologica a favore di traducenti italiani *ad hoc* più o meno felici: i puristi, di solito particolarmente attenti a spregiare tutti i forestierismi dell'informatica e di simili altre discipline pretesamente poco umanistiche, faranno bene a registrare che i due termini sono ormai dell'uso normale tra filosofi, logici e semiologi, comunità senz'altro ai loro sofisticati palati meno sgradite, ed ai loro occhi meno sospette ed invise. In questo senso, infatti, si argomentava in Barbera - Marello 2003 *i.s.* a favore in italiano di "token" e "markup" (rinunciando solo allo scomodo trattino dell'inglese *mark-up*) come prestiti non adattati (ma con le relative famiglie derivazionali adattate).

Si noti, inoltre, che *occorrenza* non può essere proposto come traducente formalmente esatto di *token* (anche se informalmente nulla ne vieta l'uso quando ciò non ingeneri particolari confusioni), dato che rende piuttosto la nozione peirceana di *instance*, che per giunta è stata da Quine 1987, p. 217 resa proprio con *occurrence*.

Dopo avere sgombrato il campo dalle "interpretazioni grammaticali" del type[44], Quine procede poi nel precisare come è che la definizione di type come classe di tutti i suoi token, che pure «fits *der* and other words well enough» (Quine 1987, p. 217) nell'esempio precedente (e nella più parte delle applicazioni della *corpus linguistics*), accada poi che «it breaks down when we press it to strings of words» (Quine cit.[45]):

> *«What about two little lines of pentameter that are fated never to get thought up? Taken as a class of their tokens, each of the lines is identically the empty class; so there is but one. This we find unaccettable. We do not want to say that every possible line of pentameter, save one, is destined someday to be uttered or written. [...]*
>
> *The postulate can be put thus:* If *a* and *b* are different strings, then the string consisting of *a* followed by *c* differs from *b* followed by *c. If types were the mere classes of their tokens, this would be false. For, if the strings* a *and* b *have actually been written but are destined never to get written with* c *appended, then the two strings with* c *appended would both be the empty class, if construed as the classes of their tokens, and would thus be identical, contrary to the postulate.*
>
> *Classes are abstract objects [...] but classes of tokens are not in general abstract enough for types. They do well enough for types of single words or signs, we saw, for we can assure the existence of tokens at that level, and thus avert emptiness of the classes. So far, so good; let us then construe types of single signs as the classes of their tokens. Types of strings of signs thereupon call for a different logical tack: we can construe them as finite* sequences *of the types of the component signs, taking 'sequence' not in its physical, spatial, or temporal sense but in its abstract mathematical sense, where failure of existence is no longer to be apprehended.»*
>
> Quine 1987, pp. 217-8.

1.4 IL MARKUP. Più complesso è definire in modo teoricamente coerente cosa sia un markup[46]. Infatti, della nozione "ingenua" sopra accennata, che è in definitiva anche quella accolta dalla iniziativa TEI (cfr. Sperberg-McQueen - Burnard 1999), sono stati più volte fatti notare i limiti. L'esposizione più recente del problema e che ci trova sostanzialmente d'accordo è quella di Buzzetti 1999, facilmente leggibile anche da non informatici. Non è questa, comunque, la sede per affrontare di petto la questione, se non almeno per rimarcare che i dieci anni intercorsi da quell'intervento hanno lasciato la questione sostanzialmente immutata. In generale, una volta esclusi dalla nozione propria di markup (cfr. sopra § 1) tutti i vari tipi di tagging, dal semplice POS-tagging (livello morfosintattico) al parsing (livello sintattico) alle più sofisticate annotazioni semantiche e testuali (che comunque non sono elementi indispensabili nella definizione di un corpus), restano pur sempre vari tipi di markup che qui giova prendere in considerazione.

Una prima distinzione possibile è tra "markup esterno", cui sono affidati i riferimenti del testo che di esso non fanno costitutivamente parte (autore, titolo, genere, capitoli, paragrafi, pagine, righe ecc.), e "markup interno e filologico", cui sono affidate le informazioni di carattere filologico (integrazioni, espunzioni, ecc.) e testuale (corsivi, prosa, verso, ecc.). Le due nozioni sono parzialmente sovrapponibili a quelle (diversamente fondate) risp. di *weakly embedded markup* 'm. (inserito in modo) sciolto' o 'non vincolato' (la trad. è di Buzzetti cit.) e di

---

[44] Cioè dalla nozione che noi chiameremmo di *lemma*: «A still further distinction can be drawn if we consult grammatical refinements. The word *der* figures as an article in its first occurrence and as a relative pronoun in its second. On this score it might be reckoned as two words, not one, even as types» (Quine 1987, p. 217).

[45] Preferisco riportare il ragionamento di Quine pressoché nella sua interezza, dato che ordinariamente la questione è sostanzialmente elusa anche nei migliori manuali di linguistica dei corpora.

[46] Come preannunciato in nota 43 e discusso in Barbera - Marello 2003 usiamo "markup" regolarmente in tondo in quanto indispensabile prestito non adattato, rinunciando a capriole ed acrobazie per trovarne un plausibile traducente italiano (perché abdicare alla internazionalità della terminologia ed alla sua accuratezza?), od allo scudo difensivo antipuristico del mantenerlo cautamente come termine straniero in corsivo. Del pari useremo liberamente le forme derivazionali adattate che se ne possono trarre (e che nei discorsi dei tecnici del settore di fatto ricorrono usualmente).

strongly embedded markup ‘m. (inserito in modo) vincolato’ (cfr. Raymond et alii 1992, pp. 3-4). Più vicina alla nozione di *embedding* è quella di *attribute* presente nella struttura CQP, dove si distingue tra *positional attributes* (riferiti ad un token, quindi *strongly embedded,* vincolati) e *structural attributes* (riferiti ad un corpus complessivamente, quindi *weakly embedded*, non vincolati). Il markup contenutisticamente esterno, e formalmente strutturale e non vincolato è spesso riferito *tout court* come “metadata”.

```
<author BrunettoLatini>
<title Tesoretto>
<genr Did>
<chapter 001>
<page 0175>
<type verse>
[...]
<s 1429>
<line 263>
Lo          lo          |art.d|          |60,0,4,6,0,0|      V   Did
Tesoro      tesoro      |n.c|            |20,0,4,6,0,0|      V   Did
conenza     cominciare  |v.m.f.ind.pr|   |111,3,0,6,0,0|     V   Did
.           stop        |punct.fi|       |70,0,0,0,0,0|      V   Did
</s>
<s 1430>
</line>
<line 264>
A           a           |adp.pre|        |56,0,0,0,0,0|      V   Did
÷l          il          |art.d|          |60,0,4,6,0,0|      V   Did
tempo       tempo       |n.c|            |20,0,4,6,0,0|      V   Did
che         che         |pd.rel|         |36,0,4;5,6;7,0,0|  V   Did
Fiorenza    firenze     |n.p|            |21,0,5,6,0,0|      V   Did
</line>
[...]
</s 1429>
</type verse>
</page 0175>
</chapter 001>
</genr Did>
</title Tesoretto>
</author BrunettoLatini>
```

Tav. 6: Attributi posizionali e strutturali nel CT (formato-CQP)

Il confine tra testo e metadata, ineludibile concettualmente e sempre tracciabile nella teoria, nella pratica è spesso confuso, perché deciso convenzionalmente, corpus per corpus, dal costruttore del corpus in base alla combinazione delle esigenze di interrogazione e delle restrizioni imposte dal software di gestione del corpus: che, nel caso del CQP[47], ad esempio, consente la interrogazione diretta dei soli attributi posizionali e non di quelli strutturali.

La necessità di distribuire i metadata tra attributi posizionali e strutturali in base alle esigenze della loro interrogabilità e non alla distinzione concettuale tra testo e markup è ben evidente da un ulteriore, più approfondito, esame delle prime righe iniziali del consueto estratto del *Tesoretto* presentato *supra* nella Tav. 6.

Qui le colonne dopo la prima rappresentano altrettanti attributi posizionali associati al token in colonna uno (colonna 2. *lemma*; 3. e 4. *POS tag*[48], 5. *prosa/verso*, 6. *genere*); gli attributi

---

[47] Per gli analoghi ma diversi problemi della struttura TEI cfr. Buzzetti 1999.

[48] Per la distinzione tra il POS-tag tipato (HDF) del primo capo ed il tag morfologico non tipato (MSF), preceduto dal “tag HDF numerico collassato”, del secondo campo, cfr. in questo volume Barbera ¶ 8.

strutturali XML sono costruiti nella prima colonna intorno al testo (*weakly embedded)*: “autore” <author BrunettoLatini>, “titolo” <title Tesoretto>, “genere” <genr Did>); od a porzioni di esso (*strongly embedded*): “capitolo” <chapter 001>, “pagina” <page 0175>, “riga” <line 263>, “frase” <s 1429>, “tipo” <type verse>). Questa distribuzione, evidentemente, rispecchia la ripartizione concettuale tra testo e markup solo in modo imperfetto: il “genere letterario” ed il “tipo prosa/verso” (propriamente markup) sono rappresentati, oltre che da attributi strutturali, anche da due attributi posizionali in modo da potere effettuare query miste; le qualifiche di “capitolo” (propriamente testo) e quelle di “pagina” e “riga” (propriamente markup), pur evidentemente diverse, sono entrambe attributi strutturali ugualmente *strongly embedded*; i POS-tag ed il confine di “frase” non sono markup, ma sono marcati del pari come attributi strutturali *strongly embedded*.

L’esempio di Tav. 6 evidenzia anche bene che, se è opportuno distinguere, almeno teoricamente, markup da tagging, di fatto non c’è (ossia: il software non consente / richiede) una vera distinzione formale tra markup interno e tagging, come si può ancora meglio vedere dall’esempio[49] seguente (Tav. 7), tratto dalla parte iniziale del solito frammento del *Tesoretto* nel CT.

Qui sono state aggiunte anche tre colonne per il markup filologico (colonna 8. “msform”, la lezione dei manoscritti; 9. “philform”, la rappresentazione filologica; 6. “s/n” la differenza o meno del token da quanto alle colonne 8 e 9) e tre per il tagging delle multiword (colonna 10. “lemma-MW”; 11. “POS-tag-MW”, 12. numero di “costituente-MW”). Oltre che con i tre suddetti attributi posizionali, i confini delle multiword sono anche demarcati da un tag XML di markup interno (attributi strutturali *strongly embedded*), <mwl>, ricorrendo, dunque, ad una strategia formalmente mista.

In definitiva, cosa in un corpus sia markup (e che tipo di markup) e cosa sia tagging, è più l’architettura del software usato (ed i suoi effetti sulle strategie di interrogazione) a deciderlo, che non ragioni teoriche e concettuali.

[49] Come già nella tavola 5 < | > sostituisce per ragioni di spazio l’accapo

```
<author BrunettoLatini> | <title Tesoretto> | <genr Did> | </chapter> | <chapter 001> | </page> | <page 0175> | </type> | <type verse> | [...] | </s> | <s 1429> | </line>
| <line 263>
Lo          lo           |art.d|          |60,0,4,6,0,0|         V  n  Did  Lo        Lo        --                --              0
Tesoro      tesoro       |n.c|            |20,0,4,6,0,0|         V  n  Did  Tesoro    Tesoro    --                --              0
conenza     cominciare   |v.m.f.ind.pr|   |111,3,0,6,0,0|        V  n  Did  conenza   conenza   --                --              0
.           stop         |punct.fi|       |70,0,0,0,0,0|         V  n  Did  .         .         --                --              0
</s> | <s 1430> | </line> | <line 264>
<mwl>
A           a            |adp.pre|        |56,0,0,0,0,0|         V  n  Did  A         A         a°÷l°tempo°che°   51,0,0,0,0,13   1
÷l          il           |art.d|          |60,0,4,6,0,0|         V  n  Did  ÷l        ÷l        a°÷l°tempo°che°   51,0,0,0,0,13   2
tempo       tempo        |n.c|            |20,0,4,6,0,0|         V  n  Did  tempo     tempo     a°÷l°tempo°che°   51,0,0,0,0,13   3
che         che          |pd.rel|         |36,0,4;5,6;7,0,0|     V  n  Did  che       che       a°÷l°tempo°che°   51,0,0,0,0,13   4
</mwl>
Fiorenza    firenze      |n.p|            |21,0,5,6,0,0|         V  n  Did  Fiorenza  Fiorenza  --                --              0
</line> | <line 265>
froria      fiorire      |v.m.f.ind.ipf|  |112,3,0,6,0,0|        V  n  Did  froria    froria    --                --              0
,           comma        |punct.nfi|      |71,0,0,0,0,0|         V  n  Did  ,         ,         --                --              0
e           e            |conj.co|        |50,0,0,0,0,0|         V  n  Did  e         e         --                --              0
<mwl>
fece        fare/-si/    |v.m.f.ind.pt|   |113,3,0,6,0,0|        V  n  Did  fece      fece      fare°frutto°      113,3,0,6,0,12  1
frutto      frutto       |n.c|            |20,0,4,6,0,0|         V  n  Did  frutto    frutto    fare°frutto°      113,3,0,6,0,12  2
</mwl>
,           comma        |punct.nfi|      |71,0,0,0,0,0|         V  n  Did  ,         ,         --                --              0
</line> | <line 266>
<mwl>
sì          sì           |adv.gn|         |45,0,0,0,8,0|         V  n  Did  sì        sì        sì°che°           51,0,0,0,0,12   1
ch'         che          |conj.sb|        |51,0,0,0,0,0|         V  n  Did  ch'       ch'       sì°che°           51,0,0,0,0,12   2
</mwl>
ell'        ella         |pd.per.s.no|    |37,3,5,6,0,0|         V  n  Did  ell'      ell'      --                --              0
era         essere       |v.a.f.ind.ipf|  |212,3,0,6,0,0|        V  n  Did  era       era       --                --              0
de          di           |adp.pre|        |56,0,0,0,0,0|         V  n  Did  de        de        --                --              0
÷l          il           |art.d|          |60,0,4,6,0,0|         V  n  Did  ÷l        ÷l        --                --              0
tutto       tutto        |pd.ind|         |32,0,4,6,0,0|         V  n  Did  tutto     tutto     --                --              0
```

Tav. 7: Testo markuppato, tokenizzato, lemmatizzato e taggato anche per MW e categorie filologiche (formato-CQP finale)

Questo è tanto più vero in presenza di metadata molto ricchi e complessi, articolati parte in ampie header XML di attributi strutturali, parte in attributi strutturali *strongly embedded* nel testo, e parte in attributi posizionali opportunamente studiati per rendere interrogabili determinate informazioni, come ad es. nel corpus VALICO.

```
<project id="1" charset="ansi" format="txt" date="iso_code8601">
        <corpus name="valico" version="1.0" class="learner" language="iso_code-639_2|639_1"
                content="text" contenttype="written">
        <corpussize texts=num.testi token=num.token types=num.type average_textsize=200>
        <corpussource adress="università torino" country="italia" date="iso_code8601"
                contact="Elisa Corino"
        </corpus></corpussource></corpussize>
  <doc idN="num." data="iso_code8601">
      <HEAD tipo_forma=" " tipo_produzione=" " test=" ">
        <gruppo nome="articolo" num=2 num_totale=" "></gruppo>
        <origine_testo luogo="toponimo" paese="isocode 3166-1"  ist="tiposcuola"
                ist_nome="nomescuola" topics="" keywords=""></origine_testo>
        <testo esecuzione="ms" qualita="origFC" cap-min=""></testo>
        <trascr> <\trascr>
        <autore specifiche="f" eta_min=19 eta_max=25 status="2" annualita="+">
            <lingua1 1="croato" 2="" 3="" 4="" 5=""><\lingua1>
            <lingue L2="inglese" L3="italiano" L4="tedesco" L5=""><\lingue>
            <contatto_lingua scolarizzazione="sp" permanenza=24 permanenza_luogo="Verona"
                esposizione="sc|.|."> </contatto_lingua>
        </autore>
      <\HEAD>
      <BODY>
      [ testo;
      attr. pos.: word, POS, lemma,
      atrr. strutt. :  CORR, INS, VAR, LAC, CORR  ]
      </BODY>
      <ref>
        <stel>nomecognome_F.txt,nomecognome_T.txt,titolo_G.txt,0</stel>
        <cons ="stazione_C.txt"</cons>
        <txtext>0</txtext>
        <imgext>0</imgext>
        <txtint>0</txtint>
        <imgint>0</imgint>
      </ref>
  </doc>
</project>
```

Tav. 8: Struttura del markup di VALICO (ver. Schaupp 2006)

Nella tav. 8 è compendiata la struttura sperimentale (non ancora implementata nella versione online) approntata da Annette Schaupp per il corpus VALICO (cfr. Schaupp 2006) che può dare una buona idea della complessità con cui il markup[50], si possa avvolgere intorno ed inserire dentro ad un testo, che pure ne resti idealmente distinto.

In conclusione, se non sono sempre ben definibili i diversi tipi di markup, né è meglio tracciabile il discrimine tra markup e tagging, credo si possa però asserire che la assenza o presenza di un markup qualsivoglia sia caratteristica non ambiguamente rilevabile. Ed è quanto basta ai nostri scopi.

---

[50] Rinunciamo a dare qui conto di ogni elemento di questo complesso schema, rimandando al volume in corso di stampa Corino - Marello *i.s.*

Ma da solo non basta ad individuare un corpus, va anche detto. Più o meno semplici header sono infatti presenti in molti tipi di documenti, come ad esempio quelli HTML (tutte le pagine web) od XML, ed in tutti i messaggi di posta elettronica o post ad un newsgroup. Se consideriamo, infatti, un tipico esempio di post (Tav. 9), si noterà che il testo è preceduto da una header di metadata, e nulla più:

```
Newsgroups: es.ciencia.enologia
Subject: Re: trasiego
From: Juan Ledesma <jledesmaQUITAESTO@entelchile.net>
Date: Mon, 16 Dec 2002 18:19:58 -0400
Message-ID: <3DFE518E.6090603@entelchile.net>
References: <3DFAD1B2.7040101@uva.es>

joscar wrote:

> Hola amantes del vino
> Tengo que hacer el trasiego de unos 25 cántaros de vino que se dignó
> darme mi pequeña viña. El asunto está en que me gustaría saber cómo
> debo lavar la cuba y con qué. He oído que una vez lavada hay que
> quemar azufre dentro. Decidme si es así o no. En caso negativo, ¿qué
> hay que hacer?
> Gracias y saludos
>

Depende del material, pero generlamente se utiliza un detergente
alcalino (a base de soda caustica) y un enjuague acido (como acido
citrico o acido peracetico), luego un enjuague y listo. Ahora si quieres
desinfectarla el acido peracetico es una buena alternativa. El quemar
azufre libera anhidrido sulfuroso que podria ayudarte a desinfectarlas,
pero de todas maneras tendrias que enjuagarlas antes de agregar el vino,
sino este podria quedar con exceso de SO2. El vapor tambien es muy util.

Salud!os
```

Tav. 9: Un tipico messaggio ("post") ad un newsgroup.

Una collezione di tali testi, senza alcun lavoro aggiunto, resterebbe solo, appunto, una collezione di testi. Se però vi implementiamo anche almeno la tokenizzazione e miglioriamo la forma del markup, giungiamo ad un formato-corpus, come si vede nella Tav. 10, che presenta il medesimo campione preparato nel formato usato per i corpora NUNC[51] (per una descrizione dettagliata del markup dei NUNC cfr. Casavecchia 2005, pp. 56-62).

[51] La rappresentazione è semplificata e compressa per esigenze di spazio.

```
<head> <doc-id>
      <idN>44</idN>
      <mess-ID><3DFE518E.6090603@entelchile.net></mess-ID>
      <mess-Ref><3DFAD1B2.7040101@uva.es></mess-Ref>
      <charset>ansi</charset>
      <lingua>spagnolo</lingua>
      <aut_NA>Juan Ledesma ,<ADDRESS@entelchile.net></aut_NA>
      <fornitore>bmanuel.org</fornitore>
      <titolo>Re: trasiego</titolo>
      <data>2002,12,16</data>
      <ora>18:19:58</ora>
      <luogo>?</luogo>
</doc-id> <set-id>
      <corpus>NUNC-ES Gneric</corpus>
      <fonte>NG</fonte>
      <f_nome>es.ciencia.enologia</f_nome>
      <f_ed>usenet</f_ed>
      <gruppo_num></gruppo_num>
      <gruppo_nome></gruppo_nome>
</set-id> <testo>
      <testoForma>post</testoForma>
      <pat>TQTQT</pat>
</testo> </head> <body>
<tit> Re : trasiego </tit>
<eLn><eLn/>
<pl> joscar wrote : </pl>
<eLn><eLn/> <qLn ind=1>
Hola amantes del vino
</qLn> <qLn ind=1>
Tengo que hacer el trasiego de unos 25 cántaros de vino que se dignó
</qLn> <qLn ind=1>
darme mi pequeña viña . El asunto está en que me gustaría saber cómo
</qLn> <qLn ind=1>
debo lavar la cuba y con qué . He oído que una vez lavada hay que
</qLn> <qLn ind=1>
quemar azufre dentro . Decidme si es así o no . En caso negativo , ¿ qué
</qLn> <qLn ind=1>
hay que hacer ?
</qLn> <qLn ind=1>
Gracias y saludos
</qLn> <qLn ind=1></qLn> <eLn><eLn/> <tLn>
Depende del material , pero generlamente se utiliza un detergente
</tLn> <tLn>
alcalino ( a base de soda caustica ) y un enjuague acido ( como acido
</tLn> <tLn>
citrico o acido peracetico ) , luego un enjuague y listo . Ahora si quieres
</tLn> <tLn>
desinfectarla el acido peracetico es una buena alternativa . El quemar
</tLn> <tLn>
azufre libera anhidrido sulfuroso que podria ayudarte a desinfectarlas ,
</tLn> <tLn>
pero de todas maneras tendrias que enjuagarlas antes de agregar el vino ,
</tLn> <tLn>
sino este podria quedar con exceso de SO2. El vapor tambien es muy util .
</tLn><tLn>
Salud!os
</tLn> </body>
```

Tav. 10: L’esempio di Tav. 9 in formato NUNC di base.

1.5 I CORPORA FUTURIBILI: WEB AS A CORPUS? Ad incorniciare i due concetti a nostro parere cardinali per definire un corpus (*token*, § 1.3 e *markup*, § 1.4) abbiamo simmetricamente posto la discussione di due "zone problematiche" per la nostra proposta, a seconda che venga meno il "supporto informatico", come avveniva soprattutto in passato (quindi: *corpora preistorici?*, § 1.2), o la "finitezza", come sempre più spesso oggi proposto (quindi: *corpora futuribili?* § 1.5, qui oltre); per la nozione di definitezza in sé, cfr. anche oltre, § 2.4.

Storicamente, che si arrivasse all'esplorazione delle risorse web era inevitabile: l'insufficienza quantitativa della base di dati per affrontare problematiche linguistiche specifiche sempre più complesse, ed il sempre più rapido "invecchiamento" dei materiali da considerarsi rispetto al continuo evolversi del linguaggio (anche in relazione alle nuove tecnologie ed a nuovi mezzi di comunicazione legati alla rete) non potevano che portare, negli ultimi anni, al tentativo di rendere l'intera rete Internet una sorta di mega-corpus da cui estrarre informazioni.

Ma se «the answer to the question "Is the web a corpus?" is yes» per Kilgarriff - Grefenstette 2003, p. 334 (autori che di questa rivoluzione sono stati i primi mentori e promotori), per noi la risposta andrebbe più sfumata in "perlopiù no". Le ragioni, peraltro, sono sostanzialmente simmetriche (come notato anche in precedenza): a Kilgarriff e Grefenstette importava non «the question "What is a corpus?"» ma «"What is a good corpus (for certain kinds of linguistic study)?"» (*ibidem*, p. 334); a noi invece è proprio la "semantic question" che loro accantonano come fattore di distrazione ad interessare centralmente (anche se, anzi proprio per questo, su tutto il resto siamo perfettamente d'accordo).

Le critiche sostanziali (e non essenzialmente definitorie come la nostra) più cospicue sono (prevedibilmente) giunte soprattutto dall'ala purista della disciplina: «The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present it is quite mysterious, because the search engines, through which the retrieval programs operate, are all different, none of them are comprehensive, and it is not at all clear what population is being sampled.» (Sinclair 2005). E lo stesso Sinclair ammoniva anche che: «The Web is truly bountiful, but it is important to appreciate that the idea of a corpus is much older than the Web, and it is based on "hard-copy" concepts, rather than cyber-objects like web "pages". A corpus expects documents (including transcripts) to be discrete, text to be linear and separable from non-text, and it expects documents to fall into recognisable sizings, similar to hard-copy documents. A normal corpus has no provision for hypertext, far less flashing text and animations. Hence all these familiar features of the Web are lost unless special provision is made to retain them» (Sinclair 2005a).

In realtà è utile anzitutto notare come l'etichetta "Web as a corpus" sia oggi utilizzata in contesti differenti, come unico contenitore di almeno due situazioni distinte, più una intermedia: (1) il materiale del web reso corpus in un determinato taglio temporale, considerando le informazioni di un insieme molto ampio di testi ma comunque finito e stabile; (2) l'idea di elaborare le informazioni su materiale 'aperto', sulla rete in continuo movimento, non creando un vero e proprio corpus ma applicando ai dati *tools* di estrazione e crawling; (3) un ibrido delle due precedenti («the linguist's search engine should do periodic crawls of the Web», Lüdeling - Evert - Baroni 2006, § 3.2), paragonabile ad una collezione di *monitor corpora* (cfr. § 2.4) molto ravvicinati.

Nel caso (1) non si ha scarto alcuno dalla tradizione se non nel mezzo di procurarsi i testi, e la finitezza resta mantenuta, sicché (purché siano anche implementati tokenizzazione e markup) tali oggetti rientrano tranquillamente nella nostra definizione, come anche nella più parte delle tradizionali (non cade necessariamente il riferimento a natura linguistica, autenticità, rappresentatività, laddove formato elettronico e larghe dimensioni sono date dalla natura stessa del web). Si tratta però della soluzione più riduttiva, e che va comunque incontro a notevoli problemi legali.

È infatti soprattutto al più radicale caso (2) che si pensa quando si parla di "web as a corpus". Il pioniere di questa impostazione è stato WebCorp (cfr. Kehoe - Renouf 2002), un meta-crawler linguistico, sui cui meriti e limiti cfr. Lüdeling - Evert - Baroni 2006 § 3.1; e ad analoghe considerazioni si presta il più recente e sofisticato LSE (Linguist's Search Engine). A soluzioni di crawling diretto, dedicato e linguistico (e non appoggiato ai consueti motori di ricerca commerciali esterni come WebCorp), si muovono invece le iniziative più interessanti in corso, come quella di WaCky (***W****eb as* ***C****orpus* ***k****ool* ***y****nitiative*: cfr. Baroni - Bernardini 2006). A seconda che tutte le operazioni vengano compiute dinamicamente *on the fly* dal crawler e dalla successiva batteria di strumenti (cosa abbastanza difficile attualmente) o staticamente su grandi set di corpora transitori via via costruiti si hanno soluzioni di tipo (2) o di tipo (3).

Dal punto di vista legale, va detto, è proprio la soluzione (2) pura l'unica a non sollevare alcun problema[52], mentre le altre due, in varia misura, comportano la riproduzione di materiale comunque tutelato dal diritto d'autore. Ma (2) è anche la soluzione più difficile tecnicamente, e, proprio perché è la più innovativa, è quella che crea più problemi, tanto definitori, quanto metodologici. Problemi che però, nonostante tutto, crediamo che si possano e si debbano superare, ma che tutti derivano dalla non-finitezza.

Terminologicamente, insistere che questi nuovi strumenti che si stanno affacciando all'orizzonte non siano dei corpora ma oggetti affatto nuovi, non è un fattore negativo: anzi ne accentua la carica innovativa. Tecnicamente, infatti, le pratiche usuali della *corpus linguistics* vanno completamente rinnovate, in quanto nessuna delle operazioni statistiche[53] classiche (neanche le più semplici come il calcolo del χ^2) può infatti funzionare applicata ad insiemi non finiti. Ma i primi spunti d'applicazione, ad esempio, alla traduzione automatica (Grefenstette 1999a e Way - Gough 2003), alla lessicografia (Grefenstette 2002) od alla sintassi (Volk 2001 e 2002) sono affatto incoraggianti, e così anche, in generale, molti dei lavori raccolti in Hundt - Nesselhauf - Biewer 2006 e Baroni - Bernardini 2006.

È però dal punto di vista epistemologico che, credo, si incontra il problema più grave: il venir meno della controllabilità dovuto alla impossibilità di completa riproduzione degli esperimenti. Le necessarie condizioni *ceteris paribus*, infatti, non sono conseguibili data la costante mutevolezza della base dati. Questo farà per forza, credo, spostare l'ago della bilancia verso qualche impostazione di compromesso del tipo (3), che tenga anche conto delle istanze legali che fanno preferire le procedure "pubblicamente" *on the fly*. Ma sono problematiche che, per quanto scottanti e suggestive, non toccano il nostro orizzonte definitorio.

2. GLI ELEMENTI DELLE DEFINIZIONI TRADIZIONALI. Se nel capitolo precedente abbiamo approfondito le due caratteristiche che, combinate, secondo noi possono più utilmente essere usate come *shibboleth* nella definizione di "corpus", disegnandone anche i confini esterni (in base all'assenza-presenza dei tratti di natura informatica e finitezza), faremo ora una breve discussione dei principali elementi che sono stati finora usati dalla letteratura specialistica, come caratterizzanti l'oggetto "corpus". La disamina si basa in parte sulle "definizioni" più significative presenti in tale letteratura (che presenteremo nel § 3.1.1[54]), in parte sulla pratica stessa dei linguisti di corpora, e, più in generale, sulle loro riflessioni su di essa. La rassegna non ha ovviamente alcuna pretesa di esaustività, ma spera di essere almeno efficace e rappresentativa.

---

[52] Dell'esistenza di problemi legali avvertiva anche Sinclair 2005a: «Another tricky question is that of copyright - not the familiar copyright of publications, but the more nebulous issue of electronic copyright. In principle, under UK law, publication on the internet confers the rights on the author whether or not there is an explicit copyright statement. Every viewing of a web page on a screen includes an act of copying». Che però non si tratti di una così "nebulous issue" potrà il lettore vedere da alcuni saggi successivi in questo volume (¶¶ 5, 6 e 7).

[53] E neanche, se per quello, le tecniche di comparazione di corpora proposte dallo stesso Kilgarriff 2000a.

[54] I passi cui si riferiscono i riferimenti bibliografici, se non direttamente riportati, devono intendersi presenti nella rassegna del § 3.1, dove si troveranno anche le traduzioni eventualmente (per i criteri cfr. nota 82) fornite.

2.1 NATURA LINGUISTICA. Criterio presente (tra le definizioni riportate nel § 3.1.1) in Francis 1982, Renouf 1987, Johansson 1991, Sinclair 1996, Lewandowska - Tomaszczyk - Osborne - Schulte 2001, Blanche-Benveniste 2000, Spina 2001, Mukherjee 2002, Sampson 2004, Scherer 2006, anche se con valenze differenti.

Preliminarmente, bisogna distinguere tra natura linguistica di un corpus intesa come "basato su materiali linguistici" e come "finalizzato alla ricerca linguistica" (distinzione perlopiù lasciata implicita nella letteratura cit.).

Nel secondo caso si ha sostanzialmente un sottocaso dell' "ordinatezza finalizzata" discussa nel § 2.5. Ed è questo quello che è stato quasi sempre generalmente presupposto fin da Francis 1982 e *mutatis verbis*, Sinclair 1991, Spina 2001, Tognini-Bonelli 2001, Sampson 2004: nei termini, ossia, di Sinclair 2005, «a corpus is made for the study of language». A queste formulazioni si potrebbe opporre che quello linguistico è di solito lo scopo principale per cui si fa un corpus (non fosse che perché di solito chi lo fa è un linguista), ma non è[55] quello esclusivo. Oltre alla questione generale della multifunzionalità delle risorse, corpora specifici possono (e sono) allestiti per la ricerca letteraria (cfr. ad esempio la svolta data dalle tecniche *corpus based* nella filologia shakespeariana[56]) od anche scopi molto specifici (come i Calgary e Canterbury Corpora per il test di formati di compressione). D'altronde, anche le più antiche testimonianze di corpora (o meglio, di quelli che abbiamo qui definito come "corpora preistorici"), come il *Corpus Iuris Civilis* o il *Corpus Iuris Canonici*, certamente non erano finalizzate a ricerche linguistiche.

Nel primo caso, di corpus "basato su materiali linguistici", anche questo largamente rappresentato (cfr. ad esempio Sinclair 1987, Lewandowska-Tomaszczyk - Osborne - Schulte 2001, ecc.), il riferimento a dati linguistici risulta limitante: sarebbe meglio, in effetti, almeno predisporre una finestra più ampia, nella quale possano entrare anche materiali audiovisivi o multimediali in genere. Altrimenti progetti belli ed interessanti come il *Lancaster Corpus of Children's Project Writing* (LCCPW) rischierebbero di giocarsi l'ammissione al club dei corpora, data la forte ed importante presenza di immagini, disegni ed altri materiali che (opportunamente ed assai efficacemente) accompagnano i testi dei bimbi; ed in analoghi problemi incorrerebbero tutti i corpora più specificamente audiovisivi, categoria entro la quale sono compresi progetti diversissimi tra loro come il *Freiburg Videokorpus zur Aphasie*[57] di Peter Auer ed Angelika Bauer, ed il *Corpus lessicale audiovisivo* (LIAV) per l'analisi, la sintesi ed il riconoscimento bimodale dell'italiano parlato (cfr. Magno Caldognetto - Cosi 2002). Tutti questi "aspiranti" corpora richiedono dunque concezioni più "multimediali" di *testo* come quella proposta da Petőfi - Vitacolonna 1996 e Petőfi 2004. Ed in una nozione molto "allargata" di testo potrebbero ancora rientrare materiali radicalmente "non linguistici" ma codificati nella forma di testo scritto come le sequenze di genoma, che vengono normalmente manipolate, appunto, come corpora: la prassi è da tempo usuale nella moderna genetica molecolare (cfr. ad

[55] Od almeno non dovrebbe essere, soprattutto nell'ottica "ecologica" di economizzare e riciclare le risorse della ricerca, cfr. Barbera 2001 (e, in diversa maniera, Čermák 2002): il medesimo corpus potrebbe essere utile, ad esempio, tanto al linguista quanto allo storico della lingua od al filologo od al traduttore, ecc.

[56] In questo caso, anzi, la tradizione è ricchissima ed assai antica, ben addentro a quella che qui abbiamo chiamata la fase "preistorica" (cfr. § 1.2): «many of the advances in attribution and dating achieved by the nineteenth- and twentieth-century Shakespearians arose from their willingness to do sums», come efficacemente diceva Love 2004, p. 8c. Al di là di progetti come il *Shakespeare Dictionary Database* (Neuhaus 1988 e 1989), pensiamo soprattutto a tecniche attributive come quelle recentemente esperite da Jackson 2003 per il *Pericles*, che sarebbero certo facilitate avendo a disposizione veri corpora anziché relativamente semplici (per quanto già utilissime!) collezioni di testi digitalizzati come quelle della LION (Chadwick-Healey Literature Online).

[57] Che Lemnitzer - Zinsmeisters 2006, pp. 124-5, descrive come «Videoaufnahme zur Familieninteraktion mit Aphasikern; Longitudinalstudien, in der 10 Familien über ein Zeitraum von einem Jahr nach Entlassung des Aphasikers aus der Klinik beobachtet wurden», e che presenta «Trascription, Digitaliesierung, Alignierung von Text mit Videospur».

es. Ficket - Guigó 1993[58]) ed anzi esistono software dedicati per fare ciò, come il CodonCode Aligner, un noto “Software for DNA Sequencing”[59].

Un terzo punto, legato ad entrambi i precedenti ma in realtà presupposto del secondo, esplicitamente sollevato dalla sola (che ci risulti) Tognini-Bonelli 2001, p. 3, è il principio che un corpus sia “*langue*-oriented”: l’idea, assai interessante, che “corpus evidence yield insights into langue” (contrapposto ad un testo, inteso come atto di *parole*) è probabilmente presente a tutti i linguisti dei corpora[60], almeno a quelli (come noi) di antica coscienza saussuriana. È forse però così basilare da restare di solito implicita; a meno che, naturalmente, non si rilegga in questi termini la polemica, ormai classica[61], tra «intuition-based and observation-based grammars» (per riprendere un succoso articolo di Aarts 1991). Per quanto ricca teoricamente sia questa idea (ed importante per il dibattito circum-generativo su competenza ed intuizione da un lato, ed esecuzione e corpora dall’altro), non è però probabilmente molto raccoglibile in contesto definitorio, in quanto difficilmente discriminante, e comunque legata ad una questione più teorica che formale.

2.2 AUTENTICITÀ. Che un corpus sia interessante in quanto collezione di dati autentici, è idea cui è accordato particolare rilievo in numerose definizioni: Sinclair 1987 e 1991, Biber et alii 1998, Sampson 2004, Rossini Favretti 2000a, Renouf 1987, Lewandowska - Tomaszczyk - Osborne - Schulte 2001, Tognini-Bonelli 2001, Bowker - Pearson 2002, Granger - Hung - Petch-Tyson 2002, Hunston 2002, Mukherjee 2002, McEnery 2003, Mitkov 2003a, Granger 2004, Scherer 2006, McEnery - Costelatos 2006.

Principio guida dell’intera disciplina fin dai suoi albori, e prima (si veda ad es. l’esperienza di Fries 1952 valorizzata da Sampson 2004a p. 9; cfr. qui *supra* § 1.2), è in effetti l’attenzione prestata alla raccolta di dati reali, estratti da una lingua effettivamente prodotta dai parlanti, in polemica con ogni esempio studiato a tavolino da una linguistica “introspettiva”; e, comprensibilmente, tale elemento è pertanto maggiormente in rilievo nelle definizioni che puntano sull’autonomia della “linguistica empirica” teoricamente definita da Sampson (e cfr. l’ampia discussione in Tognini-Bonelli 2001, p. 55-57).

Storicamente, proprio su tale elemento si snodò infatti il dibattito contro l’intuizionismo generativo e le note obiezioni chomskiane sull’inadeguatezza dei corpora a rappresentare una lingua. Il rilievo dato all’autenticità dipese, peraltro, spesso anche dall’approccio “difensivo” rispetto alle affermazioni della corrente generativa. Alcuni rilievi su questo punto sono possibili (cfr. *infra*), ma nell’idea dei suoi sostenitori discriminante è in prima istanza l’esclusione di ciò che è risultato di mera intuizione del ricercatore che crei esempi a partire da processi introspettivi, un “armchair linguist” nei termini di Fillmore, che però, nell’interessante contributo del 1992, caldeggia un’auspicabile collaborazione tra *corpus* ed *armchair linguistics*. Questa è l’osservazione su cui tornano più volte anche McEnery - Gabrielatos in uno dei contributi più recenti in materia (2006): «A point that all writers defining corpus linguistics agree upon is that corpus linguistics is empirical, in that it examines, and draws conclusions from, attested language use, rather than intuitions» (p. 34). Gli autori ne argomentano l’importanza anche in relazione alla presenza del contesto, indispensabile per l’interpretazione dei dati (*ib.*, p. 109): «data collected introspectively is decontextualized” e “what might sound awkward and ungrammatical out of context can become quite grammatical in context» (*ib.*, p. 98). Anche

[58] Le cui “tecniche” da biologo non a caso suonano abbastanza comprensibili anche al linguista di corpora!
[59] Usato dai biologi ma non a caso pubblicizzato anche su siti di *corpus linguistics* come Athel.com.
[60] «A corpus essentially tells us what language is like, and the main argument in favour of using corpus is that it is a more reliable guide to language use than native speaker intuition is» (Hunston 2002, p. 3).
[61] «It is not easy nowadays to recall how idiosyncratic, in the years immediately after the LOB Corpus was completed in 1978, was Geoffrey Leech’s assumption that a good way to discover how the English language works is to look at real-life examples» (Sampson 1996, p. 14).

nelle "Raccomandazioni" EAGLES di Sinclair 1996, p. 7 «the default value for Quality is authentic. All the material is gathered from the genuine communications of people going about their normal business. Anything which involves the linguist beyond the minimum disruption required to acquire the data is reason for declaring a special corpus».

Si tratta, a nostro parere, del caso più evidente di analogia con la situazione che denunciava l'epigrafe di Rosen (e cfr. più diffusamente tutto il passo: Rosen 1972, pp. 30-3) per la "forma sonata": l'impostazione teorica cui si ispirano i costruttori di corpora (là era la poetica e le creazioni dei costruttori classici di sonate) che detta il modello formale e definitorio cui conformarsi. Tutto normale, naturalmente: ma è opportuno, crediamo, cercare di tenere distinti gli elementi astratti, formali e definitori dell'oggetto-corpus da quelli teorici e programmatici della linguistica che tali oggetti-strumento usa.

Si possono, comunque, trovare notevoli eccezioni a questo criterio, legate tutte alla relativa ambiguità della nozione di *autenticità*.

La prima è data dai corpora di comunicazione uomo-macchina, come ad es. ADAM, il corpus di dialoghi annotati per interfacce vocaliche avanzate di Claudia Soria e Vito Pirrelli, che comprende anche dialoghi uomo-macchina. La parte spettante alla macchina non è "autentica" nel senso di "naturally occurring language", ed in questa direzione è possibile accettare anche l'idea di non-autenticità dei dati per una lingua non prodotta spontaneamente dai parlanti ma frutto interamente di dispositivi informatici, come ad esempio in sintesi vocale, generazione automatica di testi[62], traduzione automatica. Oostdijk 1991, invece, esplicitamente considerava dati autentici solo la lingua prodotta spontaneamente, distinguendo nettamente le «potential utterances or utterances that originate from experiments in a laboratory environment». Questa posizione ci pare certo troppo riduttiva, poiché porta ad escludere, oltre ai corpora puramente sintetici cui accennavamo sopra, anche una vasta gamma di corpora elicitati quali ad esempio corpora basati su interviste guidate o *learner corpora* costruiti con esercizio per elicitare un determinato lessico od una sintassi particolare; non siamo, si ha l'impressione, qui in presenza dell'eterno conflitto tra *corpus linguistics* e generativismo, tra lingua reale e lingua costruita *ad hoc*, ma ci troviamo piuttosto di fronte ad un'esasperazione del concetto di lingua autentica che costringe la nozione di corpus in un'area fin troppo limitata. Le "Raccomandazioni" EAGLES di Sinclair 1996, infatti, non escludevano la produzione "guidata", imponendone però di dichiararne le specifiche e le particolarità per distinguere lo "special corpus" così creato da quelli "tradizionali" contenenti produzioni spontanee. Si noti, tra l'altro, che solo alcuni di questi corpora "sintetici' sono anche "non finalizzati a ricerche linguistiche" (e che quindi ricadono anche sotto il quanto al § 2.1), ma assolutamente non tutti.

La seconda eccezione, sempre legata all'elasticità della nozione di "autenticità", riguarda non la "sinteticità" dei testi ma la loro "genuinità", ossia testi che variamente si pongono sotto l'insegna della riscrittura della copia o del plagio: esemplare di questa tipologia è il METER Corpus, che si propone di servire da *training corpus* per il riconoscimento automatico e misurazione del riuso testuale in àmbito giornalistico[63].

È quindi evidente la necessità di un uso perlomeno cauto del criterio dell'autenticità.

---

[62] Di largo utilizzo, ad esempio, per il test di architetture di biblioteche digitali XML sono i corpora XML sintetici, generati da software automatici come ToXgene o X007 Benchmark.

[63] «The corpus consists of a set of news stories written by the Press Association (PA), the major UK news agency, and a set of stories about the same news events as published in nine various British newspapers. In some cases the newspaper stories are rewritten from the PA source; in other cases they have been independently written by the newspapers' own journalists.» (METER 2002).

2.3 RAPPRESENTATIVITÀ. Criterio presente in Francis 1982, Sinclair 1987, 1991 e 1996, Biber et alii 1998, Rossini Favretti 2000, McEnery - Wilson 2001[64], Lewandowska-Tomaszczyk - Osborne - Schulte 2001, Spina 2001, Tognini-Bonelli 2001, Mukherjee 2002, Sampson 2004, Sinclair 2005, Scherer 2006, Baker - Hardie - McEnery 2006, Lemnitzer - Zinsmeister 2006, McEnery - Costelatos 2006, Lüdeling - Kytö - McEnery 2006, Jones - Tschirner 2006.

McEnery - Wilson 2001 la pongono come prima *main heading* della loro articolata definizione nella versione "sampling and representativeness". Anche Lemnitzer - Zinsmeister 2006 la considerano qualifica discriminante: «Das erste Kriterium [repräsentative Korpora] qualifiziert Korpora als solche und unterscheidet sie von anderen Sammlungen linguistischer Daten». È evidente che, mirando all'analisi induttiva di dati linguistici autentici per risalire a conclusioni valide ad un livello più ampio e generalizzato dello studio linguistico[65], la base empirica debba necessariamente aderire a criteri di rappresentatività, costituisca cioè un campione, un "sample" della lingua analizzata che ne riproduca idealmente, seppur "in miniatura", tutte le caratteristiche (sulle orme del concetto di *parole* specchio della *langue*, cfr. § 2.1), pur nell'impossibilità di avere, in ultima analisi, le stesse identiche caratteristiche della lingua oggetto di analisi. Questa aporia è stata ben espressa da Sinclair 2005[66], ma è stata talvolta portata alle estreme conseguenze come in Wikipedia 2007ru, dove si sostiene recisamente che «не существует объективного критерия отбора текстов для корпусов. Каждая группа руководствуется своей логикой…»[67]. In effetti, «the principle of representativeness, crucial as it is, has been used and referred to rather loosely and vaguely in both corpus and non-corpus linguistics, and the differences between existing suggest that there are differing views on how the general concept translates into the size [cfr. infatti § 2.7!] and structure of a large, versatile corpus» (Kučera 2002, p. 246).

«So we sample, like all the other scholars who study unlimitable phenomena», diceva Sinclair 2005 cit.: ma tuttavia non è sempre chiaro cosa si intenda con il termine "sample". Se gli autori citati poc'anzi corredano in effetti le loro definizioni con una descrizione di cosa si intende per "campione di lingua" e quali caratteristiche esso deve avere per soddisfare il principio della rappresentatività (od almeno in relazione a cosa esso debba essere considerato), altri autori (Aarts 1991, CIC 2006) si limitano a sottintendere nelle loro definizioni il criterio di rappresentatività, accontentandosi della sola espressione "sample": «a corpus is understood to be a collection of samples of *running text*» (Aarts1991); «a corpus is a large collection of samples of a language» (CIC 2006); e così anche Jones - Tschirner 2006 parlano di un "rational sample", senza però chiarirne il significato.

Il problema della rappresentatività diventa comunque molto spesso quello del "bilanciamento" (per la esplicitazione concreta e teorica di questa nozione è fondante Biber 1993[68], che

---

[64] In linea di massima limitiamo qui i riferimenti a McEnery - Wilson 2001, ad esclusione di McEnery - Wilson 1996 e 2007, a meno di significative differenze.

[65] Garside - Leech - McEnery 1997 vedono la finalità principale del corpus nell'essere «designed to represent a particular language or language variety».

[66] «Everyone seems to accept that no limits can be placed on a natural language, as to the size of its vocabulary, the range of its meaningful structures, the variety of its realisations and the evolutionary processes within it and outside it that cause it to develop continuously. Therefore no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself. Fine. So we sample, like all the other scholars who study unlimitable phenomena. We remain, as they do, aware that the corpus may not capture all the patterns of the language, nor represent them in precisely the correct proportions. In fact there are no such things as "correct proportions" of components of an unlimited population. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen» (Sinclair 2005).

[67] 'Non c'è nessun criterio oggettivo per la selezione di testi per il corpus. Ogni gruppo richiede la sua logica...'

[68] Se ne veda anche la recente ripresa da parte di Meyer - Nelson 2006, che tra l'altro ricordano (p. 107) il "cyclical process" proposto da Biber 1993, che vedrebbe, dopo l'identificazione dei tipi di testo da includere, la

però non entra in sede definitoria). Una discussione particolarmente efficace, che mette in luce i vari problemi della nozione, è data da Tognini-Bonelli 2001, p. 55-57, che si pone direttamente il problema di come e cosa campionare ("sampling") per conseguire tale rappresentatività, *ibidem* pp. 59-62.

Alla rappresentatività è strettamente legato anche il dibattito sulle dimensioni di un corpus (più il corpus è ampio, più lo si potrebbe ritenere rappresentativo di una lingua): questione per cui cfr. infra § 2.7.

E se rappresentatività è bilanciamento, si tratta anche, per dirla alla Engwall (1994), di una questione di "choice e non di chance", in cui la fase di progettazione e finalizzazione non è dunque secondaria: l'ipotesi della scelta casuale implica infatti che la varietà dei testi, qualunque sia la sua estensione, non sarà rappresentativa dell'intera popolazione, né di uno specifico gruppo; la selezione programmata ed organizzata dei materiali, invece, deve aderire alla definizione della lingua che il corpus si propone di rappresentare ed i confini entro i quali le scelte avvengono devono essere ben tracciati[69]. Ma per il problema della finalizzazione cfr. oltre, § 2.5.

Rappresentatività (§ 2.3), finalizzazione (§ 2.5) e dimensione (§ 2.7) sono pertanto parametri interrelati: è infatti chiaro però che l'attenzione per la rappresentatività di un corpus significa attenzione per la funzione cui il corpus stesso è destinato; ed ai fini di ricerche specifiche, una raccolta di dimensioni limitate ma in sé equilibrata, ben etichettata ed accuratamente verificata, comprendente in sè la dimensione linguistica che interessa studiare, risulterà più efficace di una raccolta ampia ma non proporzionata al suo interno.

Nonostante la sua asserita centralità, la rappresentatività non è in genere ben a fuoco nelle definizioni: da una parte c'è chi (come McEnery - Wilson 2001) pone la questione al centro eppure non fa riferimento a criteri formali e ad un dato grado di pianificazione, dall'altra c'è un'ampia schiera di definizioni (tra gli altri Atkins - Clear - Ostler 1992, Blanche-Benveniste 2000, Bowker - Pearson 2002, MNSz 2005…) che fa invece esplicita menzione dell'autenticità dei materiali e (talvolta) della presenza di criteri per la loro raccolta, ma poi non nomina altrettanto esplicitamente la rappresentatività medesima quale elemento fondante della definizione, celandola forse implicitamente (cfr. ad es. Bowker - Pearson 2002) in un limbo situato tra quelle che qui chiamiamo "ordinatezza finalizzata" (cfr. § 2.5) ed "autenticità" (cfr. § 2.2). Biber et alii 1998 invece segnano la differenza ed individuano nella raccolta finalizzata ciò che distingue un corpus da una semplice raccolta di testi, ed in particolare sarebbe la rappresentatività a segnare la cifra della distanza tra i due. Anche Lewandowska - Tomaszczyk - Osborne - Schulte 2001 sottolineano la non casualità della raccolta («the concept of corpus does *not* cover any arbitrary collection of language data»), a differenza, ad esempio, di Aarts 1991, Blanche-Benveniste 2000, Meyer 2002, CIC 2006 e Kolde 2006, che invece non ne specificano le caratteristiche, propendendo anzi per l'identità tra corpus ed «any collection of texts» (Meyer 2002) «of any length» (Aarts 1991) che «can come from anywhere» (CIC 2006).

La differenza rispetto ad un semplice archivio od a una raccolta di testi su formato elettronico, questione già emersa nella discussione precedente, è specificamente illustrata da Atkins - Clear - Ostler 1992 (cfr. anche § 2.5): un archivio sarebbe una raccolta di testi in formato elettronico non connessi tra loro (ad es. l'Oxford Text Archive); una libreria elettronica di testi (in

---

costruzione di un piccolo corpus pilota che ne testi empiricamente la copertura in termini di variabilità linguistica, indicando eventuali modifiche da apportare in un ciclo di costruzione quasi continuo.

[69] Bilanciare un corpus significa anche considerare la diversa lunghezza dei testi che lo compongono, questione che interseca la trattazione in 2.7, su cui Sinclair 2005 critica una consumata consuetudine: "There is no virtue from a linguistic point of view in selecting samples all of the same size. [...] The integrity and representativeness of complete artefacts is far more important than the difficulty of reconciling texts of different dimensions. Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size".

inglese *Electronic text library* od ETL, in francese *textotheque*) si avvicinerebbe invece al concetto di corpus nelle sue caratteristiche di standardizzazione e controllo dei contenuti, le sue possibilità di interrogazione sono però più limitate (tali sarebbero ad esempio la Banca dati testuale dell'OVI o la *repository* di Semanticsarchive); un corpus infine sarebbe costruito secondo criteri di selezione precisi, studiati in modo da raggiungere gli scopi di analisi che ci si è proposti. Questo criterio, per quanto suggestivo (ed utile appare soprattutto la distinzione tra ETL ed archivi), non crediamo sia però da solo efficace a discriminare tra corpus e non-corpus, basandosi su una opposizione graduale e non privativa, e quindi troppo dipendente dal *judicium* di chi la usa[70]. Gli autori, inoltre, menzionano il fatto che per ragioni di brevità la parola *corpus* è solitamente sovraestesa a tutte e tre le tipologie di raccolta. A noi pare invece, alla luce di ciò che abbiamo discusso precedentemente, che l'ipergeneralizzazione sia fuori luogo e che invece il termine debba venire circoscritto unicamente alle raccolte che possiedono tutte le caratteristiche qui escusse.

Un ultimo interessante sviluppo è quello di Lemnitzer - Zinsmeister 2006, che discutono la questione della rappresentatività e generalizzabilità dei dati anche in termini statistici: è difficile raggiungere una rappresentatività "pura" in relazione ad una *Grundgesamtheit* che è in continua crescita. Nessun corpus può a rigore rendere conto di una lingua in modo esaustivo, se questa si accresce ogni ora di nuove frasi e testi: «[…] ist ein Corpus immer nur eine Art Stichprobe, von der wir nicht wissen, ob sie wirklich repräsentativ ist und die Verhältnisse so widerspiegelt, wie sie auch in der Gesamtheit sind» (p. 54), da cui la proposta di confronto e verifica del medesimo fenomeno linguistico analizzato su più corpora dalle caratteristiche diverse; ed un'altra possibile conseguenza di ciò potrebbe essere la messa in discussione, in varie maniere (*monitor corpora*, *web corpora*, ecc.) del concetto di "finitezza", per cui cfr. il paragrafo seguente.

2.4 FINITEZZA. Invocata esplicitamente, a quanto ci risulta, in quasi nessuna definizione, la ritroviamo soltanto in Lüdeling - Kytö - McEnery 2006 e McEnery - Wilson 2001, in questi ultimi come seconda "main heading" (pp. 30-31, "a *finite-sized* body of machine-readable text"), senza ulteriori specificazioni.

In realtà la natura finita dei corpora è probabilmente assunzione data per scontata piuttosto che volontariamente elusa. Se consideriamo, infatti, l'uso della statistica come una caratteristica individuante da sempre la *corpus linguistics* rispetto ad altre discipline linguistiche (fin dai suoi prodromi friesiani), è condizione matematicamente banale che gli insiemi di elementi su cui opera debbano essere finiti. Più in generale, inoltre, la finitezza di un corpus ne garantisce la possibilità di operare entro confini scientificamente ed univocamente stabiliti dal linguista, non solo a livello di bilanciamento del materiale in esso contenuto (che non potrebbe essere tenuto sotto controllo in un corpus "aperto"), ma anche a livello di completa ripetibilità, *ceteris paribus*, degli esperimenti.

Va da sè che questo reca notevoli problemi ad una delle tendenze più all'avanguardia nella moderna linguistica dei corpora: il fenomeno dei cosiddetti *web corpora*[71]. E come questo problema possa essere affrontato lo abbiamo visto nel § 1.5.

La questione, inoltre, va anche considerata in relazione a quanto discusso in 2.3 e 2.7: da un lato la "rappresentatività" implica selezione, e quindi implicitamente finitezza; dall'altro le dimensioni "più ampie possibili" di un corpus, talvolta citate tra i criteri definitori di un corpus, portano a vagheggiare una dimensione idealmente infinita. Più realisticamente «it is hard to see why most (almost all) corpora are seen as strictly **time-limited projects** only which, when finished and having serverd their purpose, are far from being maintained, modernized, and substantially enlarged», come chiaramente indicava già Čermak 1997, p. 182. La via che si apre,

[70] Altro criterio, infatti, si proponeva qui in § 1 al fondo; e cfr. anche *infra*, § 2.5.
[71] Almeno quelli del tipo "dinamico" invocato da Kilgarriff 2001.

però, è quella dei *monitor corpora* (corpora in sé finiti, ma costruiti in serie temporale virtualmente infinita) più che dei corpora aperti (contemplabili nell'accezione più pura dei web corpora, cfr. *supra*).

Ma comunque questo tipo di rilievi (rappresentatività e dimensione infinite) non hanno seria portata sull'aspetto definitorio, rispecchiando piuttosto l'uno un corollario implicito, e l'altro una pura aspirazione ideale.

2.5 ORDINATEZZA FINALIZZATA. La caratteristica di essere ordinato ("principled" e simili in Johansson 1991, Biber et alii 1998, pp. 4, 12, Mitkov 2003a, p. 732, Tognini-Bonelli 2001, Kopotev 2003, Scherer 2006) in base ad un preciso scopo (tacitamente od esplicitamente linguistico) è spesso invocata come specifica dei corpora.

In particolare, quella d'essere costruito «according to explicit design criteria» sarebbe, a partire dalla lucida esposizione di Atkins et alii 1992, pp. 1b e 4b, ripresa ed ampliata da Kennedy 1988, l'elemento determinante per distinguere un corpus da una biblioteca di testi elettronici: «a distinction is sometimes made between corpus and a text archive or text database. Whereas a corpus designed for linguistic analysis is normally a sistematic, planned and structured compilation of text, an archive is a text repository, often huge and opportunistically collected, and normally not structured» (Kennedy 1998, p. 3).

Se strettamente inteso come subordinazione ad uno scopo linguistico, il caso ricade sotto la specifica di «finalizzato a scopi linguistici», già trattata sotto § 2.1, ed è suscettibile delle medesime eccezioni. Se più generalmente inteso in modo neutro come "uniformemente trattato" coglie invece più nel segno. L'uniformità di trattamento è senz'altro una condizione necessaria per l'esistenza di un corpus, ma non sufficiente (a meno che non si configuri come specificato in § 2.9 come tokenizzazione e markup).

Che, da sola, la strategia di Atkins, Kennedy, ecc. non funzioni sempre perfettamente è infatti palese: una "electronic text library" come Semanticsarchive[72] è sì "huge and opportunistically collected" (è incrementata grazie alle spontanee *submissions* degli autori), ma è assolutamente finalizzata («for exchanging papers of interest to natural language semanticists and philosophers of language», come asserito nella homepage) ed uniformemente strutturata (sono ad es. possibili anche ricerche per parola chiave, ecc.).

2.6 STANDARD. La "standard reference" è la quarta "main heading" di McEnery - Wilson 2001, p. 32, ma gli stessi ammettono di buon grado che «it is not an essential part of the definition of a corpus», anche se «there is also often a tacit understanding that a corpus constitutes a standard reference for the language variety which it represents» (*ibidem*); ed in effetti la caratteristica non è normalmente riferita nelle definizioni (si aggiungano solo Baker - Hardie - McEnery 2006, e Lüdeling - Kytö - McEnery [2006]).

Naturalmente questo è vero solo per *alcuni* corpora, tra cui soprattutto i cosiddetti "corpora nazionali"; questi saranno spesso anche i più importanti (come il Brown od il LOB corpus per l'inglese scritto, risp. statunitense o britannico), ma non sono certo i più numerosi. A fini definitori generali, quindi, non si tratta di una caratteristica rilevante.

In alcuni casi, inoltre, all'obiettivo di rappresentatività si sono problematicamente intrecciate indebite istanze normative (non è questo però, per fortuna, il caso della maggior parte dei "corpora nazionali"[73] – BNC, ČNK, ΕΘΕΓ, HNK, MNSz, NKRJa, SNK – che pure più facilmente potrebbero incorrere in tale tentazione); tale posizione è stata efficacemente rigettata

[72] Od anche la meno specifica Linguistik Online.

[73] Anzi, quando questi si configurano come "corpora letterari" possono essere articolati in più subcorpora periodo per periodo, come l'ineccepibile Eesti Kirjakeele Korpus (EKK; 'The Corpus of Estonian Literary Language').

da Sinclair 2005, che vi scorge un chiaro caso in cui «the corpus builder is adopting a prescriptive stance and is risking the vicious circle that could so easily arise, of a corpus constructed in the image of the builder».

2.7 GRANDI DIMENSIONI. È una questa caratteristica assai relativa, ma spesso invocata nelle definizioni, anche se mai da sola o raramente in termini ben definiti; Svartvik 1992a («large collections of text»), Marello 1996 («larghi insiemi di testi»); Bowker - Pearson 2002 («large collection»), Mukherjee 2002 («große, maschinenlesbare Sammlung»), McEnery 2003 («large body»), Baker - Hardie - McEnery 2006 («usually large bodies»), CIC 2006 («large collection»); notevole soprattutto quanto dice l'autoritativo (si tratta di "Raccomandazioni" EAGLES) Sinclair 1996, p. 6: «The default value of Quantity is large. A corpus is assumed to contain a large number of words. The whole point of assembling a corpus is to gather data in quantity».

Come si vede quasi tutti si limitano ad accennare ad una indefinita *große Menge* o ad un imprecisato *large body* e simili; altri però relativizzano questa nozione, come Leech 1991 («sufficiently large»), Sampson 2004, «a sizeable 'fair sample'») e Kolde 2006 («relativ große Menge»); ed anzi Jones - Tschirner 2006 giustificano le dimensioni in senso funzionale («large enough to contain a sufficient number of words to provide a useful basis»).

Naturalmente, in generale, "more data is better data" (Mercer - Church 1993, pp.18-19), «there's no data like more data» (Moore 2001), «the more data the better data» (Čermak 2002, p. 279), ecc., ma non manca anche chi minimizza: a parte la posizione estrema di Aarts 1991, p. 45 («the samples may be of any length»), si tratta sopratutto di Hunston 2002, p. 26 («Arguments about optimum corpus size tend to be academic for most people. Most corpus users simply make of as much data as is available, without worrying too much about what is not available») e, diversamente, Lemnitzer - Zinsmeister 2006, p. 105 («Man sollte sich von der Größe des Korpus nicht irritieren lassen. Letztendlich hängen Design und Größe eines Korpus von der gewählten Fragestellung ab. Für manche Fragestellungen sind sehr große Korpora unabdingbar. Man kann aber auch mit relativ kleinen Korpora Untersuchungen durchführen»). Sinclair 2005, in proposito, sostiene che «there is no maximum size», e considera specificamente due fattori nella definizione della misura *minima* di un corpus: «1. the kind of query that is anticipated from users, 2. the methodology they use to study the data» (l'interessante argomentazione è volta espressamente alla realizzazione di un corpus, ed è ovviamente meno funzionale alle esigenze dell'àmbito definitorio).

Se fosse proprio necessario definire una dimensione minima, sarebbe forse meglio ricorrere a dati oggettivi "fissi", tipo la dimensione minima utile per allenare un tagger stocastico (c. 200.000 parole, cfr. Heid 1998) o ricavare un dizionario specialistico (c. 1.000.000 di parole). Ma in realtà anche tali proposte, una volta che si considerino i corpora effettivamente esistenti, si scontrano con una ben diversa realtà: SUSANNE, ad esempio, la cui importanza, non solo per la linguistica inglese[74] ma per la linguistica dei corpora tutta, difficilmente si potrebbe sottostimare, ha "solo" 140.000 parole, ed anzi la versione addizionata di "sense annotation" SEMi-SUSANNE ne è solo una ulteriore frazione («33 documents forms [*sic*] only a small corpus, but it became the 'gold standard' I needed to evaluate my word-sense disambiguation algorithms» Powell 2006); siamo pertanto ben al di sotto della supposta "soglia minima", la cui ragione d'essere sarà pertanto da mettere fortemente in dubbio.

Inoltre, su teorizzazione ed uso di corpora di piccole dimensioni soprattutto per la glottodidattica c'è una ormai notevole tradizione di studi (cfr. ad esempio Aston 1995 e 1997, Tribble 1997, Ghadessy - Henry - Roseberry 2002), ed una consolidata pratica di cui bisogna pur tenere conto.

---

[74] «The SUSANNE scheme is so far as I am aware the first serious attempt anywhere to produce a comprehensive, fully explicit annotation scheme for English grammatical structure» (Sampson 2006).

In sostanza, non sembra che la dimensione sia in sé un riferimento utile in sede definitoria, ed anche se, come dicevamo qui nel *Decalogo* (Barbera ¶ 2), in generale «quattro testi che interroghi con la ricerca di Word non sono un Corpus, sono quattro testi», ha probabilmente ragione di scrivere de Haan 1992, p. 3 che «the conclusion seems to be that the suitability of the sample depends on the specific study that is undertaken, and that there is no such thing as the best, or optimum, sample size as such». La grandezza, cioè, va sempre relazionata alla nozione di finalizzazione, per cui cfr. supra § 2.5; e questa posizione è avallata anche da Lemnitzer - Zinsmeister 2006, che anzi aggiungono che non è neppure necessario avere i testi interi: ne bastano anche solo porzioni, purché servano agli scopi che ci si è prefissati.

2.8 FORMATO ELETTRONICO. Terza "main heading" di McEnery - Wilson 2001, è anche il criterio formale più presente (anzi, perlopiù l'unico) nelle definizioni: Renouf 1987, Svartvik 1992a, Marello 1996, Blanche-Benveniste 2000, Rossini Favretti 2000a, McEnery - Wilson 2001, Tomaszczyk - Osborne - Schulte 2001, Spina 2001, Meurman-Solin 2001, Bowker - Pearson 2002, Granger - Hung - Petch-Tyson 2002, Mukherjee 2002, Mitkov 2003a, Lemnitzer - Lobin 2004, Granger 2004, Sinclair 2005, Scherer 2006, Baker - Hardie - McEnery 2006, Lemnitzer - Zinsmeister 2006, Kolde 2006, Lüdeling - Kytö - McEnery 2006.

Il requisito, variamente espresso, della *machine-readable form*, principale discrimine dell'era storica da quella preistorica della linguistica dei corpora, come avevamo diffusamente illustrato nel § 1.2, non è enunciato nelle prime definizioni (cfr. Francis 1979 ecc.), forse nella volontà di presentarsi come continuatori della tradizione americana post-bloomfieldiana. La prima volta in cui è esplicitamente introdotto (Renouf 1987) proviene dalla scuola sinclairiana "dura e pura", e specificamente dalla premiata officina del COBUILD, sicché se ne può ben vedere la funzionalità; Sinclair tuttavia, lungi dal recidere la continuità con la tradizione precedente, ancora nel 1996 distingueva tra "corpora" e "computer corpora" in una sede importante come quella di EAGLES (Sinclair 1996, pp. 4-5).

Che quella d'essere «einheitlich kodierter elektronisch verfügbarer Textsammlungen» fosse la caratteristica principale dei moderni corpora è asserita da Lenz 2000, p. 6. A sottolineare la centralità assunta da questo fattore negli ultimi tre decenni è stata anche Spina 2001, p. 64. «Si noti peraltro – secondo scrivevamo in Barbera - Marello 2003, n. 11 – che pure i dizionari inglesi sottolineano questo aspetto della natura computerizzata del *corpus* nella moderna linguistica abbastanza tardi. Nel 1998 il *New Oxford Dictionary of English* parla esplicitamente nella prefazione di *corpus analysis* e di *evidences* trovate "using computational tools to analyse the data in the British National Corpus" e ha poi nella definizione di *corpus* il *subsense*: "a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc."». Ed anzi, la lessicografia italiana, e non solo quella, continua serenamente a trascurare l'elemento informatico (cfr. § 3.2 e sottoparagrafi).

Ma ad avere risolutamente portato in primo piano la centralità del computer nell'accezione di corpus era stata in particolare dieci anni fa Carla Marello: «C'è sempre stata una linguistica basata sullo spoglio di materiali linguistici, anche molto copiosi, ma con linguistica dei corpora, traduzione dell'inglese *corpus linguistics*, si intende oggi quella branca della linguistica che si occupa di elaborare i dati provenienti da larghi insiemi di testi immagazzinati su supporti leggibili dal computer. È dunque una linguistica dei corpora elettronici[75] [...]» (Marello 1996, p. 167).

[75] L'identificazione *tout court* della *corpus linguistics* con la *computer corpus linguistics* era già proposta in Leech 1992. Anche McEnery - Costelatos 2006, pur mirando all'interno dell'*Handbook of English Linguistics* a delineare gli impatti più interessanti dei corpora sulla linguistica inglese e non pertanto a fornire una vera e propria definizione di corpus, aprono il capitolo "English Corpus Linguistics" proprio riferendosi ad *electronic corpora* (p. 33), ribadendone l'effettiva sinonimia.

L'esplicita, contraria, dichiarazione di non rilevanza di questo fattore (al di là di una generale operazione di storiografia linguistica tesa ad individuare la tradizione empirica: cfr. § 1.2) non è frequente, ed è chiaramente espressa forse dal solo Kennedy 1998: «Historically it is not even the case that corpora are necessarily stored electronically so that they can be machine-readable, although this is nowadays the norm. […] electronic corpora can consist of whole texts or collections of whole texts. They can consist of continuous text samples taken from whole texts; they can even be made up of collections of citations» (Kennedy 1998, p. 3). In termini diversi, e più convincenti, naturalmente, vengono invece quasi sempre stabilite alterità e connessioni tra *corpus linguistics* e *computational linguistics*; distinzione scontata, certo, ma più accentuata negli autori che sottolineano la continuità con la linguistica empirica pre-informatica[76].

È stato inoltre osservato che esistono alcuni corpora di fatto "pubblicati" su supporto cartaceo, anche se «the appearence of corpora in book form is likely to remain very rare» (Mc Enery - Wilson 2001, p. 31); così, ad esempio, il Corpus of English Conversation (Svartvik - Quirk 1980; che però è semplicemente il London-Lund Corpus (LLC) nella sua *Urform*), il Corpus of Formal British English Speech (Knowles - Williams - Taylor 1996) ed i Campioni di LABLITA (Cresti 2000). È però da notare che la mera forma stampata è solo una documentazione dei corpora, che poi non sono usabili come tali se non nella loro forma elettronica, più o meno disponibile (Cresti 2000 la fornisce su CD-ROM; ed il London-Lund Corpus è facilmente reperibile).

Più rilevante era l'obiezione che "oggetti" non informatici sono tuttora usati e non sono limitati alla sola "epoca preistorica" della *corpus linguistics*: ma vi abbiamo già fatto i conti nel § 1.2 al fondo.

In tutti questi casi, comunque, il riferimento è comunque solo al supporto materiale su cui i corpora sono codificati, non al modo in cui sono codificati o vengono interrogati. Anche se non compaiono mai in contesti definitori, interessanti accenni almeno alla informaticità dell'estrazione di informazioni sono tuttavia ben presenti nella letteratura in contesti indiretti, come ad esempio in Marcus - Santorini - Marcinkievicz 1994, p. 273, quando accennano ai progressi che si possono fare in *corpus linguistics* «by investigating those phaenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora», od in McEnery - Wilson 2001, p. 17, quando scrivono che «the interest in the computer for the corpus linguist comes from the ability of the computer to carry out the processes of searching for, retrieving, sorting and calculating linguistic data».

Il riferimento al formato elettronico più propriamente *sub specie codificationis vel interrogationis*, come ad un «digital gespeicherter und für verschiedene automatische Analysen präparierter Texte» (Kolde 2006), si è però recentemente consolidato nella letteratura almeno germanica, soprattutto intorno ad un gruppo di studiosi di Tübingen (cfr. Sasaki - Witt 2004, p. 195 e Lemnitzer - Zinsmeister 2006, p. 40), ma allargato anche almeno a Genf / Genève (Kolde 2006)[77].

Ad ogni buon conto, il supporto informatico se è condizione necessaria non è però condizione sufficiente per l'esistenza di un corpus: senza altri fattori (cfr. *supra* § 1), infatti (come

[76] Cfr. ad es. Lüdeling - Kytö - McEnery [2006]: «Ever since computers were introduced in linguistic analysis, computational linguistics and corpus linguistics have been linked in three ways. In computational linguistics and corpus linguistics, techniques have been developed for structuring, annotating and searching large amounts of text. Techniques have also been designed for the qualitative and quantitative study of corpus data. In computational linguistics, corpus data are exploited to develop NLP applications».

[77] La maggiore cura posta da questi studiosi ai criteri formali anziché a quelli contenutistici, può essere dovuta proprio alla provenienza germanica della proposta, in quanto meno ossessionata di quella anglosassone (che pure è la voce certo preponderante della *corpus linguistics*) dalla tradizione della "linguistica empirica" e dei suoi difficili rapporti con la "rivale" tradizione generativa.

abbiamo più volte osservato) sarebbe impossibile tracciare il discrimine verso le raccolte di testi elettronici. Introdurre la necessità che le ricerche vengano effettuate informaticamente migliora di molto le cose, e risolve molti dei casi altrimenti ambigui, ma lascia pur sempre un certo margine di incertezza: ad esempio, nella raccolta di testi elettronici Semanticsarchive sono possibile ricerche per parole chiave; è forse questa condizione sufficiente per parlare di *corpus*? Bisogna, pertanto, introdurre anche delle restrizioni sul *modo in cui* il corpus è codificato (ossia, a nostro parere tokenizzazione e markuppatura).

Un corollario della necessità che un corpus per essere tale debba comunque essere in *machine-readable form* è il principio della "non leggibilità diretta di un corpus" (in contrapposizione ad un testo): «the essence of the corpus as against the text is that you do not observe it directly; instead you use tools of indirect observation, like query languages, concordances, collocators parsers and aligners», osservava Sinclair 2000, p. 33, cui fa eco da un punto di vista più teoretico Tognini-Bonelli 2006, p. 3. In realtà questa osservazione tiene solo per corpora grossi, non per piccoli e soprattutto letterari, dove agire puramente *corpus driven* (come ideale per Sinclair) è inverosimile: non solo la dimensione dei corpora è tale che sono scorribili manualmente, ma perdipiù contengono spesso testi che è improbabile che lo studioso non abbia già altrimenti letto, interrogando poi, di fatto, il corpus in base a ciò (in ottica quindi *corpus based*, seguendo la dicotomia sinclairiana). Chiari esempi di ciò sono il Corpus Taurinense (quale italianista non ha mai letto la *Vita nuova*?) od il Tottel's Miscellany TACT Corpus (quale anglista non ha mai letto nulla dalla *Tottel's Miscellany*?)[78].

2.9 METADATA ED ANNOTAZIONI. La presenza di un qualche tipo di markup è assai raramente menzionata nella letteratura, ed è in genere (praticamente l'unica eccezione è Sinclair 1996) limitata ai contributi più recenti. Se Baker - Hardie - McEnery 2006, p. 48 accennano solo al fatto che i corpora «usually receive some form of meta-encoding in a header», una formulazione più vincolante ed accurata si trova solo in NKRJa 2003-06, Sasaki - Witt 2004, p. 195, MNSz 2005 e Lemnitzer - Zinsmeister 2006, p. 40 a questi è da aggiungere Burnard 2005 considera come essenziale per un corpus la presenza di metadata, nel loro ruolo chiave di organizzazione dei mezzi in cui un corpus viene processato.

A parte le importanti, ma attese[79], formulazioni di Sasaki - Witt 2004 e Lemnitzer - Zinsmeister 2006, notevole è che un esplicito riferimento alla markuppatuta (sia pure limitato ai metadata bibliografici ed alle informazioni paragrafematiche) sia presente nella definizione di corpus del MNSz («nem csak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket (bekezdés, mondat)», MNSz 2005), e soprattutto del NKRJa («Разметка[80] — главная характеристика корпуса; она отличает корпус от простых коллекций (или «библиотек») текстов»[81], NKRJa 2003-06), per il quale, anzi, è il punto più importante e distintivo rispetto alla antologia di testi elettronici.

Oltre che rari, molto vaghi sono invece gli appelli all'essere un corpus «structured» (Meurman-Solin 2001, p. 6) od «einheitlich kodierter» (Lenz 2000, p. 6), che pure si possono in qualche modo ricondurre alla nozione generale di markup.

La presenza di fasce di tagging, praticamente non sempre discernibili dal markup vero e proprio, cfr. *supra* § 1.4, è pure a volte richiesta (McEnery - Wilson 2001, Sasaki - Witt 2004, Baker - Hardie - McEnery 2006, Lemnitzer - Zinsmeister 2006), ma solo come caratteristica

[78] Certo, la controbiezione che sono i testi alla base del corpus ad essere stati già letti, non il corpus medesimo, è valida; ma questo non toglie che il linguista che usa il corpus si trovi portato ad agire *corpus based* anziché *driven*.

[79] Sulla maggiore attenzione accordata agli aspetti formali dal gruppo di Tübingen (ed emanazioni), abbiamo già detto nel paragrafo precedente.

[80] Ché assumo che *газметка* sia proprio da intendere come 'markup'.

[81] Traduzioni complete dei due passi sono date al § 3.1.1, in nota risp. 83 e 84.

opzionale. In effetti non solo sono numerosi i corpora non annotati, ma vi sono anche richieste teoriche di corpora "raw" da parte della scuola sinclairiana.

Che anzi, ascetica e rigorosa come sempre, è praticamente la sola ad invocare negativamente il markup, avanzando un esplicito requisito di "semplicità" per i corpora, che è ben compendiato nelle "raccomandazioni EAGLES" di Sinclair 1996, p. 8: «The default value of Simplicity is plain text. This means that the user can expect an unbroken string of ASCII characters with any mark-up clearly identified and separable from the text. Nowadays it is likely that many texts will be in SGML format and in the future perhaps TEI. These mark-ups have been carefully designed and do not impose any additional linguistic information on the text. Largely their role in relation to text representation is to preserve in linear coding some features which would otherwise be lost. They are perceived as helpful but their presence must be recorded and the original text must be easily retrievable». Di fatto, però, i corpora considerati "default" da Sinclair oggi sono una piccola minoranza.

Accettata da molti autori è invece la possibilità di separazione dei diversi livelli di annotazione (cf. Leech 2005), anche in un'ottica di riutilizzabilità delle risorse: «Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications» (Sinclair 2005); «the raw corpus should be recoverable; the annotation should be extricable» sottolinea Leech 1997, che ricorda anche l'importanza di una documentazione accessibile agli utenti e basata per quanto possibile su analisi dei dati neutrali o 'consensuali' (*ibidem*, pp. 6-7).

Tale esigenza si può facilmente trasformare in un ulteriore requisito per un corpus, sempre di marca schiettamente sinclairiana, quello della "documentazione separata": «the default value is documented. This means that, as proposed in NERC (1994), full details about the constituents of a component are kept separately from the component itself. The model for this is the DTD or header of SGML and following that TEI. In contrast to the recommendations of those bodies corpus users seem to prefer to keep the documentation of texts in a separate place from the texts themselves and to include only a minimal header that contains a reference to the documentation» (Sinclair 1996, p. 8). Anche se è molto sensato, ed affatto condivisibile, non ci sembra comunque poter assurgere a criterio strettamente definitorio di un corpus.

Che la "uniformità di trattamento" in generale fosse una condizione necessaria ma non sufficiente lo avevamo già visto nel § 2.5; e lo stesso vale se questo trattamento si configura solo come markup, come visto nel § 1.4 al fondo; se si configura invece come tokenizzazione *e* markup, però, questa condizione diventa finalmente anche sufficiente: in altri termini, secondo noi, è forse questa la caratteristica che più consente di poter sempre tracciare un discrimine netto, non ambiguo, tra corpora e non-corpora.

3. RASSEGNA DI DEFINIZIONI RAPPRESENTATIVE. In questo capitolo presentiamo una breve rassegna di "definizioni", in varie lingue[82], parte a sostegno della trattazione analitica perpetrata nel capitolo precedente, parte come documentazione per una storia della nostra disciplina. Distingueremo (soprattutto a questo secondo fine) tra la tradizione linguistica (§ 3.1) e la lessicografica (§ 3.2).

[82] È sempre imbarazzante decidere quando dare la traduzione di una lingua straniera o no, ciò implicando presupposizioni in varia misura "impertinenti" sulle conoscenze del lettore e sul prestigio indiscutibile o meno di una lingua. Per evitare queste sgradevolezze, ci siamo attenuti ad un criterio deliberatamente non oggettivo e solo soggettivo, in quanto riposa unicamente sulle competenze degli autori: quando solo uno o due degli autori era in grado di comprendere la lingua di una citazione, ne abbiamo dato la traduzione in nota; quando tutti e tre, no. Naturalmente da lingue di cui nessuno degli autori poteva in alcun modo farsi carico, ci siamo ben astenuti dal citare. Ringraziamo inoltre Mauro Costantino, Paolo Divizia, Adriana Hanulíková, Roman Sosnowski, Irena Starčević, Ekaterina Zudina per la loro preziosa consulenza linguistica.

3.1 LE DEFINIZIONI DEI LINGUISTI. Questa rassegna di “definizioni” di corpus presenti nella letteratuea linguistica costituisce, in effetti, anche se non l'unico nutrimento, certo l’ossatura delle osservazioni presentate nel § 2; e nel suo complesso fornisce un prezioso punto di partenza per una storia della nostra disciplina.

3.1.1 GLI ESTRATTI. Tra le definizioni (presentate in ordine approssimativamente cronologico) che abbiamo compreso nella rassegna figurano solo quelle esplicite (molte di quelle indirette sono comunque discusse nei vari sottoparagrafi del § 2.). Abbiamo fatto eccezione per pochi casi di definizioni indirette (ad es. *corpus linguistics*: Marello 1996) o parziali (ad es. *learner corpora*: Granger et alii 2002, Granger 2004), solo quando particolarmente significative (soprattutto per il rilievo dato al fattore informatico).

«a sufficiently large body of naturally occuring data of the language to be investigated» (strutturalismo americano anni ’50, compendiata da Leech 1991, 2)

«a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis» (Francis 1982, p. 7 = Francis 1992, p. 17)

«a collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research» (Renouf 1987, p. 1)

«When constructing a text corpus, one seeks to make a selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypothesis can be tested» (Sinclair 1987, 2)

«The corpus, a collection of stretches of connected discourse in a single dialect, constitutes the principal source of data. The corpus is a record of performance: the utterances contained in it are unsolicited historical linguistic events and as such to be distinguished from other data, such as potential utterances or utterances that originate from experiments in a laboratory environment.» (Oostdijk 1991, p. 4)

«a corpus is a body of texts put together in a principled way, often for the purposes of linguistic research» (Johansson 1991, p. 3)

«In the Nijmegen approach, a corpus is understood to be a collection of samples of *running text*. The texts may be in spoken, written or intermediate forms, and the samples may be of any length» (Aarts 1991, p. 45)

«A collection of naturally-occurring language texts, chosen to characterize a state or variety of a language» (Sinclair 1991, p. 171)

«large collections of text available in machine-readable form» (Svartvik 1992a, p. 7)

«We distinguish four types of text collection, which we find helpful and urge the community to accept. **Archive**: a repository of readable electronic texts not linked in any coordinated way, e.g. the Oxford Text Archive. **Electronic text library** (or ETL, Fr. 'textotheque'): a collection of electronic texts in standardized format with certain conventions relating to content, etc., but without rigorous selectional constraints. **Corpus**: a subset of an ETL, built according to explicit design criteria for a specific purpose, e.g. the Corpus Revolutionnaire (Bibliotheque Beaubourg, Paris), the Cobuild Corpus, the Longman/Lancaster corpus, the Oxford Pilot corpus. **Subcorpus**: a subset of a corpus, either a static component of a complex corpus or a dynamic selection from a corpus during on-line analysis. [...] for the sake of brevity we use the word corpus to refer to all three types of collection.» (Atkins - Clear - Ostler 1992, p. 1b).

«A corpus is a body of text assembled according to explicit design criteria» (Atkins - Clear - Ostler 1992, p. 5b).

(McEnery - Wilson 1996 [1a edizione] = 2001, p. 23-24)

«con linguistica dei corpora [...] si intende oggi quella branca della linguistica che si occupa di elaborare i dati provenienti da larghi insiemi di testi immagazzinati su supporti leggibili dal computer.» (Marello 1996, p. 167).

«A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. Note that the non-committal word 'pieces' is used above, and not 'texts'. This is because of the question of sampling techniques used. If samples are to be all the same size, then they cannot all be texts. Most of them will be fragments of texts, arbitrarily detached from their contents. | A computer corpus is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.» (Sinclair [EAGLES] 1996, p. 4-5)

«a large and principled collection of natural texts» (Biber et alii 1998, pp. 4, 12)

«A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research.» (Biber et alii 1998, p. 246)

«a body of texts is a called a *corpus* - *corpus* is simply latin for 'body', and when you have several such collections of texts, you have *corpora*» (Manning - Schütze 1999, p. 6)

«Le terme de *corpus* a désigné pendant des siècles des sources documentaires caractérisées par leur exhaustivité [...]. [...] Quelles que soient les orientations, dès les premières applications aux langues vivantes, le terme de *corpus* désigne non pas simplement des collections des données de langage, mais un choix organisé de ces données.» (Blanche-Benveniste 2000, pp. 11-12)

«Una raccolta di testi, autentici e ricorrenti nell'uso, in formato elettronico, rappresentativi di uno stato o di una varietà di una lingua» (Rossini Favretti 2000a p. 41)

«[...] Korpora, einheitlich kodierter elektronisch verfügbarer Textsammlungen, [...]» (Lenz 2000, p. 6)

«Le terme *corpus* a été utilisé, de façon plus large, pour toute collection de textes rassemblée dans des bases de données informatisées. Même si la collecte n'a pas été faite de façon systématique et structurée, l'informatisation peut en faire un usage structuré» (Blanche-Benveniste 2000, p. 13-14)

«In principle any collection of more than one text can be called a corpus: the term 'corpus' is simply the Latin for 'body', hence a corpus may be defined as any body of text. It need imply [sic] nothing more. But the term 'corpus' when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under four main headings: sampling and representativeness, finite size, machine-readable form, and standard reference. […] So a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under construction. [... | ...] Corpora may exist in 2 forms: unannotated (i.e. in their existing raw rates of plain text) or annotated (i.e. enhanced with various types of linguistic information) […] » (McEnery - Wilson 2001, pp. 29, 32 e 32)

«The concept of corpus does not cover any arbitrary collection of language data. In its original older sense (cf. Latin corpus 'body'), it used to refer to any collection of writings, usually by one author. A corpus, in the sense used here, is as Leech (1991: 11) put it, a collection of machine-readable 'real-life' or, naturally occurring, linguistic data "designed or required for a particular representative function". These "databanks", as they are sometimes called, provide linguists with the materials against which they can test their hypothesis.» (Lewandowska-Tomaszczyk - Osborne - Schulte 2001, p.162)

«Raccolta strutturata di testi in formato elettronico che si assumono rappresentativi di una data lingua o di un suo sottoinsieme, mirata ad analisi di tipo linguistico» (Spina 2001)

«The concept of *corpus* is used here in the sense ‘a more or less structured compilation of digitalized texts’.» (Meurman-Solin 2001, p. 6)

«A corpus can be defined as a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis. Usually the assumption is that the language stored in a corpus is naturally-occurring, that it is gathered according to specific design criteria, with a specific purpose in mind, and with a claim to represent larger chunks of language selected according to a specific typology. Not everybody, of course, goes along with these assumptions, but in general there is consensus that a corpus deals with natural, authentic language.» (Tognini-Bonelli 2001, p. 2)

«Using a Saussurian terminology [...] text is an instance of *parole* while the patterns shown up by corpus evidence yeld insights into *langue*.

A TEXT	A CORPUS
read whole	read fragmented
read horizontally	read vertically
read for content	read for formal patterning
read as a unique event	read for repeated events
read as an individual act of will	read as a sample of social practice
instance of parole	gives insights into langue
coherent communicative event	not a coherent communicative event

The series of contrasts between corpus and text outlined above have the purpose of differentiating two sources of evidence that may appear similar but that entail very different analytical steps» (Tognini-Bonelli 2001, p. 3)

«A corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria» (Bowker - Pearson 2002, p. 9)

«a corpus will be considered a collection of texts or part of texts upon which some general linguistic analysis can be conducted» (Meyer 2002, p. xj)

«any collection of texts (or partial texts) used for purposes of general linguistic analysis»(Meyer 2002, p. xij)

«Strictly speaking, a corpus by itself can do nothing at all, being nothing other than a store of used language» (Hunston 2002, p. 3)

«Computer learner corpora are electronic collections of spoken or written texts produced by foreign or second language learners in a variety of language settings. Once computerised, these data can be analysed with linguistic software tools, from simple ones, which search, count and display, to the most advanced ones, which provide sophisticated analyses of the data.» (Granger - Hung - Petch-Tyson 2002, p. vij)

«In der modernen Korpuslinguistik versteht man unter einem Korpus eine große, maschinenlesbare Sammlung von authentischen, gesprochenen und/oder geschriebenen Texten, die als repräsentativ für den Sprachgebrauch insgesamt (bzw. für eine spezifische Gebrauchssituation) angesehen wird. Die linguistische Analyse solcher Korpora ist in ihrer typischerweise computergestützten Durchführung exhaustiv und intersubjektiv überprüfbar sowie in der Erklärung der Befunde frequenzorientiert und kontextsensitiv. Die hier in komprimierter Form angesprochenen Grundkonzepte der Korpuserstellung und der Korpusanalyse [...] sollen im folgenden systematisch charakterisiert werden.» (Mukherjee 2002, p. 47)

«A corpus (pl. *corpora*, though *corpuses* is perfectly accectable) is simply described as a large body of linguistic evidence typically composed of attested language use» (McEnery 2003, p. 449)

«A body of linguistic data, usually naturally occurring data in machine readable form, especially one that has been gathered according to some principled sampling method» (Mitkov 2003a, p. 732)

«McEnery and Wilson [1996, p. 21] (following others before them) mix the question "What is a corpus?" with "What is a good corpus (for certain kinds of linguistic study)?" muddying the simple question "Is corpus x good for task y?" with the semantic question "Is x a corpus at all?" The semantic question then becomes a distraction, all too likely to absorb energies that would otherwise be addressed to the practical one. So that the semantic question may be set aside, the definition of corpus should be broad. We define a corpus simply as "a collection of texts." If that seems too broad, the one qualification we allow relates to the domains and contexts in which the word is used rather than its denotation: A corpus is a collection of texts when considered as an object of language or literary study.» (Kilgarriff - Grefenstette 2003, p. 334)

«Представляется важным сделать еще одно, существенное, по нашему мненню, замечание. Речь идет определении самого понятия "корпус". Проблема многозначности или нечеткого употребления этого термина приводит к томы, что в некотором общем употреблении электронным корпусом называют любое собрание текстов, переденное в цифровой формат. С другой стороны, в последнее время под термином "корпус" бсе чаще понимают не просто текст (англ, "a running text"), а спецнально отовранный по тем или иным принципам языковой материал»[83] (Kopotev 2003, p. 37-8).

«A corpus, for peoples who study language and languages, is a collections of specimens of a language as used in real life, in speech or writing, selected as a sizeable 'fair sample' of the language as a whole or of some linguistic genre, and hence as a useful source of evidence for research on the language» (Sampson 2004, p. 1)

«Die definition linguistischer Korpora gestaltet sich schwierig. Im Prinzip kann ein Stapel alter Zeitungen oder eine Sammlung handschriftlicher Briefe einer bestimmten Autorin als Korpus angesehen werden. Im neuerer Zeit wird allerdings der Begriff Korpus nicht mehr in einer derartig allgemeinen Weise verstanden: Korpora werden als maschinell lesbare, digitalisierte Sprachdaten definiert. Doch auch diese Definition ist noch sehr weit gefasst. Linguistische Korpora im hier behandelten Sinne sind hauptsächlich textuelle Daten, d.h. bereits schriftlich vorliegende Texte oder transkribierte Gespräche. Sie lassen sich abgrenzen von Sammlungen linguistischer, sprachbezogener Daten, bei denen der Text nicht das zentrale Datum ist, wie z.B. Reaktionszeitmessungen in psycholinguistischen Experimenten. Audio- oder Videosignale ohne weitere Informationen fallen ebenfalls nicht unter den Begriff Korpus. Als neuer Korpusbegriff können ‚multimodale Korpora' angesehen werden, in denen Verschriftlichungen gesprochener Sprache mit anderen Modalitäten wie Gestik verbunden werden. Werden linguistische Korpora im Kontext der Texttechnologie betrachtet, ergibt sich eine weitere Verfeinerung des Korpusbegriffs. Zentral für texttechnologische Korpora sind zwei Eigenschaften: 1. Die Texte sind mit Informationen angereichert - Metainformationen oder Informationen, die die verschiedenen linguistischen Beschreibungsebenen (z.B. Morphologie, Syntax, Diskursstruktur) betreffen. 2. Die Informationsanreicherung greift auf texttechnologische Methoden zurück, also Auszeichnungssprachen [...] und Annotationskonventionen [...].» (Sasaki - Witt 2004, p. 195).

«A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.» (Sinclair 2005, p. 16).

---

[83] 'The following comment is important. We are concerned with a definition of the corpus content. There are multiple meanings or uncertain use of this term, which lead to some general tendency for the name electronic *corpus* to be given to any collection of texts put into digital format. On the other hand, recently the term corpus has increasingly been used not simply for text (English *running text*) but linguistic material especially selected on ceratin [*sic*] principles' (Kopotev 2003 ver. inglese, p. 35).

«Computer learner corpora are electronic collections of spoken or written texts produced by foreign or second language learners.» (Granger 2004, p. 124)

«A korpusz ténylegesen előforduló írott, vagy lejegyzett beszélt nyelvi adatok gyűjteménye. A szövegeket valamilyen szempont szerint válogatják és rendezik. Nem feltétlenül egész szövegeket tartalmaz, és nem csak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket (bekezdés, mondat). Az MNSz a mai magyar írott köznyelv általános célú reprezentatív korpusza kíván lenni.»[84] (MNSz 2005).

«Ein Korpus ist eine systematische Sammlung von authentischen Texten oder Textteilen. Ein Korpus bildet einen repräsentativen Ausschnitt aus einer Sprache oder Varietät ab. Korpora ermöglichen empirische Aussagen über Sprache. Die Verwendung von Korpora ist überall da sinnvoll, wo Informationen über den Sprachgebrauch benötigt werden.» (Scherer 2006, p. 15).

«Ein Korpus ist eine Sammlung von Texten oder Textteilen, die bewusst nach bestimmten sprachwissenschaftlichen Kriterien ausgewählt und geordnet werden. Unter Text sind in diesem Zusammenhang nucht nur Produkte der Schriftsprache wie Zeitungsartikel, Romane, Kochbücher, E-Mails, Briefe oder Tagebücher zu verstehen, sondern auch mündliche Äußerungen, sei es in Form von Vorträgen, Radiosendungen, Telefongesprächen oder dem zwanglosen Gespräch am Mittagstisch. Die Texte, die in einem Korpus enthalten sind, werden als Primärdaten bezeichnet. Das Korpus hat den Zweck, als Ausschnitt der Sprache zu dienen, die untersucht werden soll. Dabei ist es wichtig, sich klarzumachen, ob man eine Sprache ganz allgemein untersuchen will, also das Deutsche in seiner Gesamtheit, oder nur eine bestimmte Varietät. Unter einer Varietät versteht man eine bestimmte Ausprägung der Sprache, die durch außersprachliche Faktoren wie Zeit, Raum, Sprechergruppe oder Kommunikationssituation definiert wird. [...] Heutzutage liegen Korpora – so der Plural von Korpus – häufig in elektronischer Form vor. [...] Allerdings sind Korpora, die in reiner Papierform vorliegen, bis heute weit verbreitet.» (Scherer 2006, pp. 3-4).

«The word for corpus is Latin for body (plural corpora). In linguistics a corpus is a collection of texts (a 'body' of language) stored in an electronic database. Corpora are usually large bodies of machine-readable text containing thousands of millions of words. A corpus is different from an archive in that often (but not always) the texts have been selected so that they can be said to be representative of a particular language variety or genre, therefore acting as a standard reference. Corpora are often annotated with additional information such as part-of-speech tags or to denote prosodic features associated with speech. Individual texts within a corpus usually receive some form of meta-encoding in a header, giving information about their genre, the author, date and place of publication etc.» (Baker - Hardie - McEnery 2006, p. 48).

«A corpus is a large collection of samples of a language held on a computer. The samples can come from anywhere the language is used in speech and in writing.» (CIC 2006)

«Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte oder Äußerungensfolgen, bestehen aus den Datenselbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind. Wenn wir von *linguistischen Korpora* sprechen, dann handelt es sich um Textsammlungen mit kompletten Texten oder zumindest mit sehr großen Textausschnitten. Außerdem sollten linguistischen Korpora meist • re-

[84] «A corpus is a collection of written or spoken linguistical data. The texts are selected and classified according to certain criteria. A corpus does not necessarily contain whole texts and is not only a repository of texts: it contains their bibliographical data and marks the structural units (paragraphs, sentences). HNC wishes to be a representative general-aim corpus of present-day standard Hungarian.» (Hungarian National Corpus, *english page*: `http://corpus.nytud.hu/mnsz/index_eng.html`).

präsentativ, • durch Metadaten erschlossen und • linguistisch annotiert sein.» (Lemnitzer - Zinsmeister 2006, p. 40)

«In the first instance, a 'corpus' is simply any collection of written or spoken text. However, when the term is employed with reference to modern linguistics, it tends to imply a number of things including data in a machine-readable form, sampling, representativeness, fixed size and the idea that a corpus constitutes a standard reference for the language variety it represents.» (Lüdeling - Kytö - McEnery [2006])

«ein **(Text)korpus** ist eine relativ große Menge vorgegebener (also nicht *ad hoc* hergestellter), aus praktischen Gründen meist schriftlicher, digital gespeicherter und für verschiedene automatische Analysen präparierter Texte von gleicher oder verschiedener Textsorte bzw. Varietät aus meist einer Sprache (Ausnahme: Parallelkorpora für übersetzungswissenschaftliche Fragestellungen).» (Kolde 2006)

«[...] a corpus, i.e. a structured collection of language texts that is intended to be a rational sample of language in question. A corpus should be large enough to contain a sufficient number of words to provide a useful basis from which to work, although it has never been extablished what a threshold level should be.» (Jones - Tschirner 2006, p. 1)

«In principle, any collection of more than one text can be called a corpus, (corpus being Latin for "body", hence a corpus is any body of text). But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition.» (Mc Enery - Wilson 2007, § 3.1).

3.1.2 OSSERVAZIONI COMPLESSIVE. Poche osservazioni generali, oltre a quelle puntuali fatte nel capitolo precedente.

La complessità e specificità delle definizioni varia di molto, andando da un minimo praticamente coincidente con la definizione non-tecnica tradizionale (Manning - Schütze 1999), ad un massimo costituito da McEnery - Wilson 2001 ed Atkins - Clear - Ostler 1992; significativo, peraltro, è che anche in uno stesso autore (per di più tra i più rappresentativi: Tony McEnery), si trovino definizioni di diversa dimensione, più minimaliste o più ricche, sintomo sostanziale di una relativizzazione del problema in funzione del contesto per cui è stato formulato (come se una definizione tecnica generale di fatto non esistesse, o non fosse rilevante). Se l'esposizione più ampia è forse, come si diceva, McEnery - Wilson 2001[85], la più perspicua (almeno nella nostra ottica) è però quella di Sasaki - Witt 2004, p. 195 (cfr. *infra*).

In generale la consapevolezza della distinzione tra collezione di testi e corpus (pure a volte ben tematizzata, come in Atkins - Clear - Ostler 1992, pp. 1-2 e Tognini-Bonelli 2001) non è comunque sempre forte, come dimostrano anche scelte linguistiche generali, come ad es. quella del ΕΘΕΓ, la cui titolazione ufficiale è *Εθνικός Θησαυρός Ελληνικής Γλώσσας (ΕΘΕΓ)*, ma che sulla homepage del progetto medesimo è sottotitolato «Το Σώμα Κειμένων (Corpus) του ΙΕΛ» (ΕΘΕΓ 2006), con significativa oscillazione tra *θησαυρός* e *σώμα*.

I criteri contenutistici (con poche eccezioni) sono comunque di regola prevalenti, anzi nella storica definizione di Francis 1979 sono esclusivi, probabilmente allo scopo di rivendicare la tradizione della linguistica empirica (di cui la linguistica dei corpora dovrebbe costituire il braccio armato), o semplicemente di evidenziare la continuità con la tradizione americana post-bloomfiediana (si veda la somiglianza con la definizione "media" che ne compendiava Leech 1991, 2). Tra i requisiti esterni, formali, l'unico spesso presente è quello del formato elettronico, con alcune importanti eccezioni tedesche (soprattutto Sasaki - Witt 2004, p. 195, definizione ampia, e Lemnitzer - Zinsmeister 2006, p. 40, che è forse la migliore definizione sintetica finora proposta).

---

[85] Il pur utilissimo Atkins - Clear - Ostler 1992 è più sul versante pratico ("istruzioni per la preparazione") che non su quello descrittivo - definitorio.

3.2 LE DEFINIZIONI DEI DIZIONARI. A complemento delle posizioni espresse nella letteratura specialistica, ci sembrava interessante esaminare, anche se più cursoriamente, cosa di questa trasparisse nella pratica lessicografica, di solito comprensibilmente più tradizionalista ed attendista nel registrare i cambiamenti in atto nell'uso linguistico.

Oltre al panorama lessicografico tradizionale, interessante è parso anche ispezionare quello della "nuova" lessicografia online, tanto nella sua accezione usuale, quanto in quella enciclopedica, dove il fenomeno Wikipedia[86], il cui successo e qualità è una delle principali novità culturali del panorama contemporaneo, meritava un controllo più approfondito.

3.2.1 ESTRATTI. Presentiamo quindi, allo scopo suddetto, una campionatura ridottissima e "tagliata" in misura variabile (per ovvie ragioni di spazio), certo ben inferiore alla soglia della effettiva rappresentatività, ma sperabilmente ancora funzionale ai nostri scopi illustrativi. Oltre ai filoni italiani, inglesi e tedeschi (per diverse ragioni centrali nel nostro discorso) si è cercato anche di riportare alcuni spunti da ulteriori tradizioni linguistiche, forse più "periferiche" ma spesso nient'affatto più arretrate, come sarà presto evidente. Per la Wikipedia, si sono riportate quasi tutte le versioni più importanti, segnalando per le lingue considerate l'assenza della voce[87]; oltre a queste si sono riportati, esemplificativamente, anche estratti di un paio di "enciclopedie" linguistiche in vario modo rappresentative.

CS

«**Korpus**, -u *m* ‹l› *kniž. a odb*. **1.** celek, soupis, sbírka: *k. materiálu; statistický k.* soubor; *língv.,výp. tech.* rozsáhlý elektronicky uložených jazykových textů nebo jejich částí určený k vědeckému výzkumu jazyka: *program na vyhledávání slov v korpusu*; […]»[88] (Krause 2005, s.v., p. 446a)

[86] «Wikipedia (IPA: /ˌwiː kiː ˈpiː di.ə/) is a multilingual, Web-based, free content encyclopedia project. Wikipedia is written collaboratively by volunteers from all around the world. With rare exceptions, its articles can be edited by anyone with access to the Internet, simply by clicking the edit this page link. The name Wikipedia is a portmanteau of the words wiki (a type of collaborative website) and encyclopedia. Since its creation in 2001, Wikipedia has rapidly grown into the largest reference Web site on the Internet. [...] Wikipedia was founded as an offshoot of Nupedia, a now-abandoned project to produce a free encyclopedia. Nupedia had an elaborate system of peer review and required highly qualified contributors, but the writing of articles was seen as very slow. During 2000, Jimmy Wales, founder of Nupedia, and Larry Sanger, whom Wales had employed to work on the project, discussed various ways to supplement Nupedia with a more open, complementary project. On the evening of January 2, 2001, Sanger had a conversation over dinner with Ben Kovitz, a computer programmer, in San Diego, California. Kovitz, who was a regular on "Ward's Wiki" (the WikiWikiWeb), explained the wiki concept to Sanger. Sanger saw that a wiki would be an excellent format whereby a more open, less formal encyclopedia project could be pursued. Sanger easily persuaded Wales, who had been introduced to the wiki concept previously, to set up a wiki for Nupedia, and Nupedia's first wiki went online on January 10. There was considerable resistance on the part of Nupedia's editors and reviewers to the idea of associating Nupedia with a website in the wiki format, however, so the new project was given the name "Wikipedia" and launched on its own domain, wikipedia.com, on January 15 (now humorously called "Wikipedia Day" by some users). [...] In May 2001, the first wave of non-English Wikipedias were launched (in Catalan, Chinese, Dutch, German, Esperanto, French, Hebrew, Italian, Japanese, Portuguese, Russian, Spanish, and Swedish, soon joined by Arabic and Hungarian. [...] There are over 75,000 active contributors working on more than 5,300,000 articles in more than 100 languages. As of today, there are 1,639,121 articles in English; [...]. [...] All the text in Wikipedia, and most of the images and other content, is covered by the GNU Free Documentation License (GFDL). Contributions remain the property of their creators, while the GFDL license ensures the content will remain freely distributable and reproducible [...].» (Wikipedia 2007en).

[87] Assenza che si noti almeno anche per olandese, danese, svedese, finnico, estone, koreano, cinese, ecc.

[88] '*lett. e spec.* **1.** raccolta, collezione, unità: *corpus di materiali, corpus statistico* antologia; *ling. comp*. Ampia raccolta di testi linguistici o parti di testo immagazzinati elettronicamente, destinati all'analisi scientifica della lingua: *programma per la ricerca di parole in un corpus*; [...]'.

«Jazykový korpus je (většinou rozsáhlý) soubor textů v digitální podobě, které jsou v různé míře opatřeny metajazykovými značkami vypovídajícími o samotném textu (autor, rok vydání, žánr apod.) a zařazení jednotlivých slov do kategorie slovních druhů, o frekvenci slova v korpusu, případně dalších lingvistických a frekvenčních aspektech. Některé korpusy jsou budovány jako takzvaně vyvážené, což znamená, že by měly obsahovat vyvážený podíl textů tříděných podle žánrovosti, doby vzniku, případně dalších hledisek (mluvenost, psanost, regionálnost, užívanost apod.). K práci s korpusy se používají speciální programy, které umožňují vyhledávání slov a slovních spojení v kontextu, zjištění frekvence výskytu v korpusu i zjištění původního zdroje textu. Pro formátování textů a vkládání značek se používá zejména standardizovaného jazyka SGML, případně jeho odnože XML.» (Wikipedia, 2007cs, s.v.)[89].

DE

«**Korpus** n., pl. Korpora (lat. corpus 'Körper') **1.** Sprachl. Daten, die einer spachwiss. Analyse als Grundlagen dienen. [...] **2.** I. e. S. Sammlung einer möglichst hohen, notwendingerweise aber immer begrenzten Anzhl möglichst zusammenhängender spachl. Äußerungen (gesprochen oder/und geschrieben) aus möglichst natürl. Kommunikationssituationen. [...]» (Glück 2000, s.v., p. 384a)

«{2} **Kor|pus**, das; -, Korpora [lat. corpus = Gesamtwerk, Sammlung, eigtl. = Körper]: **1.** a) Belegsammlung von Texten od. Schriften [aus dem Mittelalter od. der Antike]; b) (Sprachw.) [als Datenbank angelegte] Sammlung einer begrenzten Anzahl von Texten, Äußerungen o.Ä. als Grundlage für sprachwissenschaftliche Untersuchungen. **2.** <heute meist: der; o.Pl.> Klangkörper besonders eines Saiteninstruments.» (Duden 2003, s.v.).

« **Korpus**[1], der; -, -se /lat./ salopp scherzh. Körper, Leib: das tut meinem K. gut. **Korpus**[2], das; -, Korpora /lat./ Wissensch. Gesamtheit von Texten, Schriften: das K. der altdeutschen Urkunden; für eine grammatische Untersuchung das K. festlegen» (*DWDS* 2003 s.v.)

«**Textkorpus**. Das *Textkorpus* (oft auch nur *Corpus* bzw. *Korpus*) ist eine Sammlung von Texten oder Äußerungen in einer Sprache, die Gegenstand einer beliebigen Darstellung oder Untersuchung wird. Eine literaturwissenschaftliche Untersuchung kann einem bestimmten Textkorpus gelten: etwa dem deutschen Roman des 20. Jahrhunderts, oder Titeln, in denen ein bestimmtes Motiv wie "Eifersuchtsmord" vorkommt. Genauso kann ein Textkorpus von Linguisten ausgewertet werden, um Regelmäßigkeiten in dieser Sprache beschreiben zu können. Eine rechtshistorische Arbeit kann ein bestimmtes Textkorpus behandeln, Gesetzestexte einer bestimmten Tradition.» (Wikipedia, 2007de, s.v.)

EN

«**corpus** [13c: from Latin *corpus* body. The plural is usually *corpora*]. (1) A collection of texts, especially if complete and self-contained: *the corpus of Anglo-Saxon verse.* (2) Plural also *corpuses.* In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analysed by means of *tagging* (the addition of identifying and classifying tags to words and

[89] 'Un corpus linguistico è una raccolta (perlopiù grande) di testi in formato elettronico, che sono in vari modi marcati metalinguisticamente per ogni testo (autore, anno di edizione, genere, ecc.) e classificati in ogni parola per parti del discorso, per frequenza della parola nel corpus, eventualmente in altri aspetti linguistici o statistici. Alcuni corpora sono costruiti come (così si dice) bilanciati, il che significa che dovrebbero abbracciare parti bilanciate di testi classificate in base a genere, epoca di composizione, eventualmente altre angolature (parlato, scritto, regionale, uso, ecc.). Propriamente coi corpora si usano particolari programmi, che consentono di ricercare le parole e le espressioni nel contesto, trovare la frequenza delle occorrenze nel corpus ed anche delle originali fonti del testo. Per la formattazione dei testi e l'inserzione di marche si usa principalmente il linguaggio standardizzato SGML, eventualmente la sua variante XML.'

other formations) and the use of concordancing programs. *Corpus linguistics* studies data in any such corpus [...]. (T.McA)» (*OCEL* 1992 s.v.)

«A representative sample of language, compiled for the purpose of linguistic analysis, is known as a *corpus*.» (Crystal 1997, p. 414b)

«**cor·pus** (kôr'pəs) *n.*, *pl.* **-po·ra** (-pər-ə), abbr. **cor.** **1.** A large collection of writings of a specific kind or on a specific subject. **2.** The principal or capital, as distinguished from the interest or income, as of a found or estate. **3.** *Anatomy*. **a.** The main part of a bodily structure or organ. **b.** A distinct bodily mass or organ having a specific function. **4.** *Music*. The overall length of a violin. [...]» (*AHD* s.v., p. 421b)

«**corpus** /'kô:'pəs/ ► noun (pl. *corpora* or *corpuses*) **1** a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject: *the Darwinian corpus*. ■ a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research. **2** Anatomy the main body or mass of a structure. ■ the central part of the stomach, between the fundus and the antrum. - ORIGIN late Middle English (denoting a human or animal body): from Latin, literally 'body'. Sense 1 dates from the early 18th cent.» (*OED* 2005, s,v.)

«**cor·pus** (kôr'pəs) *n.*, *pl.* -po·ra (-pər-ə). **1.** A large collection of writings of a specific kind or on a specific subject. **2.** A collection of writings or recorded remarks used for linguistic analysis. **3.** *Economics*. a. The capital or principal amount, as of an estate or trust. b. The principal of a bond. **4.** *Anatomy*. a. The main part of a bodily structure or organ. b. A distinct bodily mass or organ having a specific function. **5.** *Music*. The overall length of a violin. [Middle English, from Latin.]» (*TFD* 2007 = Answers 2007, s.vv.)

«In linguistics, a **corpus** (plural *corpora*) or **text corpus** is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis, checking occurrences or validating linguistic rules on specific universe. [...]» (Wikipedia, 2007en, s.v.)

ES

«**corpus**[2]. (De or. lat.). 1. m. Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación.» (*DRAE,* s.v.)

«**Corpus lingüístico**. Un *Corpus lingüístico* es un conjunto, normalmente muy amplio, de ejemplos reales de uso de una lengua. Estos ejemplos pueden ser textos (típicamente), o muestras orales (normalmente transcritas). [...]» (Wikipedia, 2007es, s.v.)

FR

«**corpus** [kɔʀpys] *n. m.* (1863; "hostie", 1642; mot lat. "corpus"). ♦ 1° *Dr.* Recueil de pièces, de documents concernant une même discipline. [...] ♦ 2° *Ling*. Ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.» (*PR*, s.v., p. 396b)

«**corpus** subst. masc. **A.** – *PHILOL., SC. HUM.* Recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc. [...]. – *LING.* Ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique. *Le corpus des textes parus d'un journal, d'une revue; un corpus littéraire; le corpus du vocabulaire français.* **B.** – *Spécialement* **1.** *DR.* Recueil, collection du droit romain. *Le corpus juris,* p. abrév., *le corpus.* **2.** *ÉLECTRON., INFORM.* Ensemble de données exploitables dans une expérience d'analyse ou de recherche automatique d'informations. *Perforation de corpus.* **Rem.** On rencontre le composé *sous-corpus.* Partie d'un corpus. [...].» (*TLFi*, s.v.)

[fr.wikipedia ha solo «une ébauche à compléter concernant la littérature»].

HR

«**kórpus m.** […] **5.** *lingv.* skup svih pojavnica koji se izrađuje računalnim putem za uporabu u konkordancijama, tezarijima i sl. [*Hrvatski nacionalni* ~]» (Jojić 2002, s.v.)[90].

[hr.wikipedia manca della voce].

IT

«**còrpus**, sm. Latin. Raccolta completa di leggi, di norme giuridiche (che interessano un dato settore); complesso organico di scritti concernenti una determinata materia. [...] 2. Locuz. lat. *Habeas corpus*. [....]» (*GDLI* s.v., III. p.812.c.)

«**corpus** (còr-pus) s.neutro lat., in it. *s.m.* 1. Raccolta ordinata e completa di opere o di autori [...]. 2. Campione prelevato a fini scientifici dal linguista [...]» (*DOLI* s.v., p. 706a)

«**corpus** /'kɔrpus/ s. neutro lat. (pl. *corpora*); in it. s.m. inv. (meno freq. pl. orig.) – ♦ 1. Raccolta completa di testi e di opere costituita secondo un particolare criterio SIN corpo [...]. ♦ 2. ling. Raccolta di brani, singoli enunciati o altri dati linguistici, che vengono analizzati per definire la struttura di un sistema linguistico SIN campione – ET [...] nell'accez. 2. entrato dall'ingl. [...] a. 1969 (2)» (*DISC* s.v., p. 617b)

[it.wikipedia manca della voce].

JP

«コーパス (cor･pus /kps｜k－/ (cor･po･ra/kp()r｜k－/) ラテン語｢身体，総体｣の意1. A (文書などの)集成，集積，全集. B (資料の)総体，集成資料〔of〕. 2 (人 動物の)死体. (新英和中辞典 第6版 （研究社）).»[91] (Kenkyusha 2004)

«コーパス（corpus;「身体」を意味するラテン語に由来、複数形はcorpora（こーぽら）だが通常使われない）とは、電子化された自然言語の文章から成る巨大なテキストデータである。言語学や自然言語処理などの研究に使うため、言語的な情報（品詞、統語構造など）が付与されていることが多い。元となる文章を集めるにあたり著作権などの法的問題が発生する他、電子化の手間などが発生するため、大規模なコーパスの作成には相当の費用と時間がかかる。現在日本では国立国語研究所が一億語の収録を目指すKOTONOHA計画をすすめている。»[92] (Wikipedia 2006ja, s.v. コーパス)

MA

«**korpusz** fn **1.** *Tud* Vmely kérdés(kör)re vonatk. írások, források összesége. | Adattár, rendezett adathalmaz. **2.** *Vall Műv* Kereszten Krisztus testének szoborszerű abrázolása. **3.** *Zene* Hangszekrény. [lat]»[93] (Pusztai 2003, s.v.).

[hu.wikipedia manca della voce]

---

[90] 'Insieme di parole [propriamente *pojavnica* 'tutte le forme che compaiono di una parola, token'] elaborate in formato elettronico per l'uso nelle concordanze, tesauri [~ *nazionale croato*]'.

[91] 'cor･pus [...] Dal latino: corpo, insieme. 1a raccolta, insieme (di scritti o simili). b insieme (di dati) raccolta di dati. 2 cadavere (di persona o di animale)'.

[92] 'Corpus: (parola latina che significa CORPO, (pl. CORPORA generalmente non usato)) grande raccolta di dati testuali di una lingua naturale convertita in formato elettronico. Viene molto utilizzato nelle ricerche linguistiche e di natural language processing per il rilevamento di dati lnguistici (parti del discorso, costruzioni sintattiche, ecc.) Oltre ai problemi di tipo legale nel raccogliere scritti, a causa del lavoro di conversione dei dati in formato elettronico, la costituzione di un corpus su larga scala impegna un ammontare considerevole di risorse e tempo. Attualmente in Giappone l'Istituto Nazionale per la Lingua Giapponese (National Institute for Japanese Language) continua il progetto KOTONOHA puntando a raggiungere milioni di termini'.

[93] 'sost. 1. *Scient.* raccolta di scritti, fonti concernenti un argomento | Base di dati, grande quantità di dati ordinati. 2. *Art. rel.* Statua raffigurante il corpo di Cristo crocifisso. 3. *Mus*. Cassa di risonanza. [lat.]'

PL

«**korpus** ‹łac. *corpus* ‘ciało’›. **1.** *książk* **a)** […]. **b)** zasadnicza część, podstawa czegoś; zrąb. ○ Korpus serii wydawniczej. ○ Korpus tekstów, dzieł Żeromskiego.»[94] (Dubisz 2004, s.v.)

«**Korpus (językoznawstwo)** - zbiór tekstów służący badaniom lingwistycznym, np. określaniu częstości występowania form wyrazowych, konstrukcji składniowych, kontekstów w jakich pojawiają się dane wyrazy. Nowszym zastosowaniem korpusów jest uczenie maszynowe w przetwarzaniu języków naturalnych.» (Wikipedia, 2006pl, s.v.)[95]

PT

«**Corpus** s. m. compilação de documentos ou informações relativos a uma disciplina ou tema; LINGUÍSTICA conjunto finito de enunciados representativos de uma determinada estrutura (Do lat. *Corpus*, “corpo, conjunto, matéria”)» (*DLP*, s.v.,)

[pt.wikipedia manca della voce]

RO

«**Córpus** s. N. 1. culegere, colecție de date, de texte, inscripții, legi. 2. garmond. (<lat., fr. corpus)» (FLORIN 2004, s.v.)

[ro.wikipedia manca della voce]

RU

«**Лингвистический корпус**. Лингвистическим корпусом называют собрание текстов, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой. Иногда корпусом («корпус первого порядка») называют просто любое собрание текстов, объединённых каким-то общим признаком (языком, жанром, автором, периодом создания текстов).» (Wikipedia, 2006ru, s.v. *Корпусная лингвистика*)[96]

SK

«**korpus** -u M. **1**. KNIŽ. telo (VÝZN. 1): *kríž s korpusom* - **2**. podstatná časť niečoho: *korpus skrine, korpus hudobného nástroja, korpus torty* - **3.** ODB. súbor skúmaných prvkov: *štatistický korpus; textový korpus* súbor textov v počítačovom spracovaní určený na vedecký výskum»[97] (SLEX 1999, s.v.)

«**Korpus (jazykoveda)**. Korpus textov v jazykovede je ohraničený súbor jazykových výpovedí zaznamenaných písmom alebo na zvukovom nosiči, ktorý spracováva na vedecko-výskumné a učebné ciele; množina textov používaných na lingvistický opis a argumentáciu; v užšom zmysle elektronická databáza jazykových prvkov spolu s prostriedkami efektívneho vyhľadávania.»[98] (Wikipedia 2007sk, s.v.)

---

[94] ‘‹Lat. *corpus* ‘corpo’›. 1. *Lett.* a) Parte fondamentale, base di qualcosa, fondamento. Corpus della collana editoriale. Corpus dei testi, delle opere di Żeromski’

[95] ‘Raccolta di testi che serve nelle ricerche linguistiche, ad esempio per la definizione delle frequenze d’occorrenza di forme morfologiche, di costruzioni sintattiche, dei contesti in cui compaiono le parole. Tra i nuovi usi dei corpora c’è l’insegnamento ai software nel trattamento automatico delle lingue naturali’.

[96] ‘Per corpus linguistico si intende una collezione di testi, marcati secondo una regola precisa e dotati di un motore di ricerca specializzato. Talvolta per corpus si intende semplicemente una qualsiasi raccolta di testi, uniti da qualche caratteristica comune (lingua, genere, autore, periodo di composizione dei testi)’.

[97] ‘1. LETT. corpo (ACC. 1): *una croce col corpo* - 2 elemento sostanziale di qualsiasi cosa: *corpo del mobile, corpo dello strumento musicale, corpo della torta* - 3 SPEC. raccolta di elementi da analizzare: *corpus statistico, corpus di testi*; raccolta di testi elaborati al computer, destinati ad analisi scientifiche’.

[98] ‘Corpus (Linguistica). Un corpus di testi in linguistica è una raccolta delimitata di espressioni linguistiche, immagazzinata per iscritto o su base audio ed utilizzata per analisi scientifiche o scopi didattici; raccolta di testi utilizzata per la descrizione ed argomentazione linguistica, in senso stretto banca dati elettronica di unità linguistiche con efficaci possibilità di interrogazione’.

3.2.2 OSSERVAZIONI COMPLESSIVE. Il panorama, va detto, non è certo dei più confortanti (e ne avevamo già avuto sentore dai pochi assaggi ammaniti nella *Premessa* di questo articolo), in quanto l'accezione specialistica del termine è spesso mancante, elusa o non messa appropriatamente a fuoco[99]. E per l'italiano, in particolare, questo è purtroppo verissimo: la più maschia figura ve la fa ancora (come spesso avviene) il Sabatini - Coletti, che pure manca il discrimine informatico; mentre il Battaglia addiritura scambia definizione generale con accezione specialistica (giuridica) in un quantomeno inopportuno *pot-pourri* definitorio. Comunque se Atene piange, Sparta non ride: la situazione lessicografica nelle altre lingue europee "maggiori" in genere non è perlopiù rosea. Curiosa e degna di nota è, ad esempio, la voce del *Trésor de la langue française*, che, oltre che portare una accezione specialistica (quella filologica) a definizione generale, distingue poi da questa una accezione "informatica", senza che la nostra linguistica rientri propriamente né nell'una né nell'altra.

Il formato elettronico, pur raramente invocato, è comunque l'unica caratteristica formale (talvolta) menzionata, almeno nei dizionari "tradizionali", e, come avevamo notato in precedenza (§ 2.8), «pure i dizionari inglesi sottolineano questo aspetto della natura computerizzata del *corpus* nella moderna linguistica abbastanza tardi[100]» e solo «nel 1998 il *New Oxford Dictionary of English* parla esplicitamente nella prefazione di *corpus analysis* e di *evidences* trovate "using computational tools to analyse the data in the British National Corpus" e ha poi nella definizione di *corpus* il *subsense*: "a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc."» (Barbera - Marello 2003, n. 11). Tale "apparizione" è peraltro preceduta da quella nel *Companion* del 1992 (*OCEL* 1992), che è anche stata utilizzata come punto di partenza da qualche linguista (la presenza più rilevante è quella di Ball 1997).

In generale, comunque, anche estrapolando in modo mirato solo alcune accezioni, e concentrandosi solo su opere lessicografiche di recente edizione, è evidente già dalla pur brevissima panoramica di definizioni una diffusa inconsapevolezza di ciò che concerne la *corpus linguistics*. "Corpus" è infatti considerato nella maggior parte dei casi (quando una accezione "umanistica" vi è!) nel senso che originariamente diede vita al "nostro" odierno concetto di *electronic corpus,* ovvero la semplice nozione di raccolta di dati o collezione di testi, con prevalente riferimento ad antologie letterario-filologiche o raccolte giuridiche; gli elementi di rilevanza sono assai pochi: il riferimento a "representativeness" e "tagging" nella definizione inglese (*OCEL*), «lo más extenso y ordenado posible» nella definizione spagnola (*DRAE*), «rendezett adathalmaz» in quella ungherese (Pusztai 2003), rimandano ad alcune peculiarità di strutturazione interna già discusse ma oltre non si va, e sono più numerosi gli elementi mancanti se non addirittura fuorvianti (cfr. il riferimento a *database*). Il problema non risiede evidentemente nelle controversie definitorie oggetto del dibattito che abbiamo prima esemplificato, ma, molto probabilmente in una lacuna a monte che la linguistica dei corpora non ha ancora saputo superare in termini divulgativi e che questo contributo vuole almeno invitare a colmare.

La scena cambia però significativamente se si abbandona il campo della lessicografia tradizionale e ci si sposta su quella online, allargando l'indagine anche alle altre lingue europee di cultura, oltre alle canoniche, ed anzi estendendo i sondaggi al di là dell'Occidente stesso. Teoricamente, bisognerebbe distinguere tra impostazione effettivamente lessicografica e più propriamente enciclopedica: i semplici dizionari online (che abbiamo rappresentato con TheFreeDic-

---

[99] L'insufficienza della lessicografia in materia era lamentata già da Francis 1982, p. 7: «The Random House Dictionary of the English Language (1967) gives as its definition of *corpus*: '*Ling*. a body of utterances or sentences assumed to be representative of and used for grammatical analysis of a given language or dialect.' This is essentially the same as definition 3b in Webster's New International. But it is too restricted». E le cose, evidentemente, non sono cambiate molto fino ad ora.

[100] La riluttanza all'aggiornamento è evidente anche nella esemplificazione portata dal *TLF* «perforation de corpus", quando sono ormai trent'anni che in informatica non si "perfora" più.

tionary ed Answers.com) non si diversificano in nulla da quelli tradizionali (anzi, riproducono l'*American Heritage*), ma lo stacco qualitativo è netto con la enciclopedica Wikipedia. Che il divario qualitativo sia semplicemente da attribuire alla diversità di base delle due impostazioni è però, crediamo, sostanzialmente da escludere: una piccola campionatura di "enciclopedie" tradizionali basta a fare constatare che non si distaccano molto dalla media della corrispondente tradizione lessicografica (allo scopo abbiamo riportato in § 3.2.1 due buoni ed affatto diversi esempi di "enciclopedia linguistica", Crystal 1997 e Glück 2000, l'uno inglese e lontanissimo dall'impostazione lessicografica, non essendo neppure alfabetizzato, e l'altro tedesco e più del tipo del lessico specialistico).

In genere, comunque, per ogni tradizione linguistica, la definizione delle Wikipedia è sempre migliore di quella del corripondente dizionario standard tradizionale (si confronti ad esempio la definizione della Wiki polacca con quella di Dubisz 2004). Notevole è anche come le voci più interessanti vengano più dalla periferia che dal centro (imperniato sulla roccaforte dell'inglese), soprattutto nella situazione web: i dizionari (anglofoni) "medi" più diffusi (entrambi basati sull'*AHD*) sono molto scarsi, e poco migliore è la voce di Wikipedia inglese; la voce di Wikipedia giapponese, invece, è in assoluto la migliore tra le campionate (accenna al formato elettronico, dà dettagli linguistici e computazionali, e menziona persino il problema legale!).

4. CONCLUSIONI E DEFINIZIONE. Una definizione più accurata di quella, preliminare, proposta in § 0, che tenga ossia conto, oltre che (a) delle esigenze, a nostro avviso essenziali, di adeguatezza e demarcazione, anche (b) delle caratteristiche più tradizionalmente espresse dalla tradizione della linguistica dei corpora, potrebbe pertanto essere la seguente, in cui alle condizioni (a) soddisfano restrizioni formali necessarie, ed alle condizioni (b) restrizioni contenutistiche facoltative:

> *Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi.*

Si noti che tale definizione è assai liberale circa i criteri e scopi con cui un corpus è allestito (ammettendo anche corpora di finalità non linguistiche) e riguardo ai materiali a partire dai quali è costituito (ammettendo anche corpora non tradizionalmente testuali[101], o composti per campionatura[102]), ma molto restrittiva sulle condizioni formali dell'oggetto medesimo. L'idea infatti era proprio quella di seguire le due "raccomandazioni" espresse nelle epigrafi, depurando (*à la* Rosen) la definizione dalle caratteristiche storiche di cui è stata gravata, e riducendola (*à la* Wittgenstein) alla sua funzione puramente architettonica. Questo ha comportato sceverare il corpus *sui propri generis* da un lato dagli strumenti non informatizzati e dalle (variamente costituite ed efficienti) collezioni di testi (cfr. § 1.1), e dall'altro dai molteplici "oggetti" che stanno nascendo dall'uso del web (cfr. § 1.5): l'uno il passato ed il presente, l'altro senz'altro il futuro. Il che naturalmente nulla pregiudica circa la possibilità (particolarmente auspicabile soprattutto per il secondo caso) che un linguista dei corpora usi o sviluppi anche questi altri strumenti oltre

---

[101] Cfr. la concezione più "multimediale" di *testo* in Petőfi - Vitacolonna 1996 e Petőfi 2004.

[102] La possibilità di ciò era stata, assai correttamente, presa in considerazione praticamente dal solo Sinclair 1996, p. 4, che usava «the non-committal word 'pieces' [...] and not 'texts'. This is because of the question of sampling techniques used. If samples are to be all the same size, then they cannot all be texts. Most of them will be fragments of texts, arbitrarily detached from their contents»; meno esplicitamente è però presente anche nella definizione di corpus del MNSz («Nem feltétlenül egész szövegeket tartalmaz», MNSz 2005).

ai corpora popriamente detti (o che linguisti precedenti l'era informatica abbiano compiuto operazioni epistemologicamente analoghe): anzi, alcuni dei saggi presenti in questa silloge (in modo diverso Korzen ¶ 12 e Conte ¶ 22) documentano anche questa possibilità; ma almeno i corpora veri e propri che egli usa saranno ora definiti esplicitamente in base a caratteristiche formali inequivocabili, con indubbio vantaggio vuoi dal punto di vista teorico, che da quello pratico (si pensi ad es. all'aspetto legale, come vedremo nei tre prossimi contributi).

BIBLIOGRAFIA.

AARTS
1991 Jan Aarts, *Intuition-based and Observation-based Grammars*, in AIJMER - ALTEMBERG 1991, pp. 44-62.

AARTS - MCMAHON
2006 *The Handbook of English Linguistics*, edited by Bas Aarts and April McMahon, Malden MA (USA) - Oxford (UK) - Carlton (Australia), Blackwell Publishing, 2006 "Blackwell handbooks in linguistics".

AA. VV.
1992 *The American Heritage Dictionary of the English Language*, Boston - New York, Houghton Mifflin Company, 1992_3.
2004 *Trésor de la langue française informatisé*, disponibile online a `http://atilf.atilf.fr/tlf.htm`.

AHD → AA. VV. 1992.

AIJMER - ALTENBERG
1991 *English Corpus Linguistics. Studies in Honour of Jan Svartvik,* edited by Karin Aijmer and Bengt Altenberg, London - New York, Longman, 1991.

AINSWORTH - BISBY
1995 *[Geoffrey Clough] Ainsworth & [Guy Richard] Bisby's Dictionary of Fungi*, by D[avid] L. Hawksworth, P[aul] M. Kirk, B[rian] C[harles] Sutton and D[avid] N. Pegler, 8th edition prepared by the International Mycological Institute, Wallingford (UK), CAB International, 1995.

ALLORA - BARBERA
¶ 5 Adriano Allora - Manuel Barbera, *Il problema legale dei corpora. Prime approssimazioni*, in questo volume, pp. 109-118.

ALMEIDA COSTA - SAMPAIO E MELO
1999 Joaquim Almeida Costa - António Sampaio e Melo, *Dicionário da língua portuguesa*, 8. edição revista e actualizad, Porto, Porto Editora, 1999_8 [1952_1].

ANSWERS
2007 *Corpus*, pagina online: `http://www.answers.com/topic/corpus`, s.d. [last checked 15 February 2007].

ARMSTRONG
1994 *Using Large Corpora*, edited by Susan Armstrong, Cambridge (Mass.) - London (En.), The MIT Pres, 1994 "A Bradford Book", "ACL-MIT Press Series in Computational Linguistics" [= "Computational Linguistics" XIX $(1993)^{1-2}$].

ASTON
1995 Guy Aston, *Corpora in Language Pedagogy: Matching Theory and Practice*, in Cook - Seidlhofer1995, pp. 257-270.
1997 Guy Aston, *Small and Large Corpora in Language Learning*, in LEWANDOWSKA-TOMASZCZYK - MELIA 1997, pp. 51-62; disponibile online alla pagina `http://www.sslmit.unibo.it/~guy/wudj1.htm`.

ATKINS - CLEAR - OSTLER
1992 [Beryl T.] Sue Atkins - Jeremy Clear - Nicholas Ostler, *Corpus Design Criteria*, in "Literary and Linguistic Computing" VII (1992)[1] 1-16.

ATKINS - ZAMPOLLI
1994 *Computational Approaches to the Lexicon*, edited by B[eryl] T. S[ue] Atkins and A[ntonio] Zampolli, Oxford - New York, Oxford University Press, 1994.

BAKER - HARDIE - MCENERY
2006 Paul Baker - Andrew Hardie - Tony McEnery, *A Glossary of Corpus Linguistics,* Edinburgh, Edinburgh University Press, 2006.

BALL
1997 Catherine N. Ball, *Tutorial: Concordances and Corpora*, Georgetown University, 1997, online a `http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html`.

BARBERA
2000 Manuel Barbera, *La linguistica dei corpora sul Web. Un'introduzione*, conferenza tenuta presso l'università di Trieste il 2 maggio 2000. Inedita.
2001 Manuel Barbera, *From EAGLES to CT Tagging: a Case for Re-usability of Resources*, in RAYSON et alii 2001, pp. 40-44.
¶ 2 Manuel Barbera, *Il decalogo della Corpus linguistics. (Tanto* Esodo *20,2-17 e* Deut. *5,6-21 erano diversi)*, in questo volume, pp. 21-23.
¶ 8 Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in questo volume, pp. 135-168.
¶ 23 Manuel Barbera, *Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra* guidelines *e prolegomeni*, in questo volume, pp. 373-388.

BARBERA - MARELLO
2003 *i.s.* Manuel Barbera - Carla Marello, *Corpo a corpo con l'inglese della* corpus linguistics*, anzi, della linguistica dei corpora*, in *Atti del Convegno Internazionale Lingua italiana e scienze, Firenze, Accademia della Crusca 6-8 febbraio 2003*, in corso di stampa.

BARONI - BERNARDINI
2006 *WaCky! Working Papers on the Web as Corpus*, edited by Marco Baroni and Silvia Bernardini, Bologna, GEDIT edizioni, 2006, disponibile online alla pagina `http://wackybook.sslmit.unibo.it/`.

BATTAGLIA [- BÀRBERI SQUAROTTI]
1961-2004 Salvatore Battaglia [† 1971 - Giorgio Bàrberi Squarotti dal vol. IV], *Grande dizionario della lingua italiana*, Torino, U.T.E.T., 1961-.... ¶ **I.** *A-Balb*, 1961; **II.** *Balc-Cerr*, 1962; **III.** *Cert-Dag*, 1964; **IV.** *Dah-Duu*, 1966; **V.** *E-Fin*, 1968; **VI.** *Fio-Grau*, 1970; **VII.** *Grav-Ing*, 1972; **VIII.** *Ini-Libb*, 1975; **IX.** *Libe-Med*, 1975; **X.** *Mee-Moti*, 1978; **XI.** *Moto-Orac*, 1981; **XII.** *Orad-Pere*, 1984; **XIII.** *Perf-Po*, 1986; **XIV.** *Pra-Py*, 1988; **XV.** *Q-Ria*, 1990; **XVI.** *Rib-Roba*, 1992; **XVII.** *Robb-Schi*, 1994; **XVIII.** *Scho-Sik*, 1996; **XIX.** *Sil-Sque*, 1998; **XX.** *Squi-Tog*, 2000; **XXI.** *Toi-Z*, 2002;

Supplemento a cura di Edoardo Sanguineti, 2004; *Indice degli autori citati* a cura di Giovanni Ronco, 2004.

BLANCHE-BENVENISTE
2000 Claire Blanche-Benveniste, *Types de corpus. Introduction*, in BILGER 2000, pp. 11-15.

BERGMAN - PAAVOLA
2003 *The Commens Dictionary of Peirce's Terms. Peirce's Terminology in His Own Words*, edited by Mats Bergman & Sami Paavola, `http://www.helsinki.fi/science/commens/dictionary.html`.

BIBER
1993 Douglas Biber, *Representativeness in Corpus Design*, in "Literary and Linguistic Computing" VIII (1993) 243-57; già presentato al *Pisa Workshop on Textual Corpora 1992*; poi riedito in SAMPSON - MCCARTHY 2004, pp. 174-197.

BIBER et alii
1998 Douglas Biber - Susan Conrad - Randi Reppen - Jean Aitchison, *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge, Cambridge, 1998 "Cambridge Approaches to Linguistics".

BILGER
2000 *Corpus. Méthodologie et applications linguistiques*, edité par Mireille Bilger, Paris, Honoré Champion Éditeur - Presses Universitaires de Pepignan, 2000 "Bibliothèque de l'INaLF.Les français parlés" 3.

BONFANTINI et alii
1980 Charles Sanders Peirce, *Semiotica*, Testi scelti e introdotti da Massimo A. Bonfantini, Letizia Grassi, Roberto Grazia, Torino, Einaudi, 1980 "Paperbacks" 115.

BOWKER - PEARSON
2002 Lynne Bowker - Jennifer Pearson, *Working with Specialized Language. A pratical Guide to Using Corpora*, London - New York, Routledhe, 2002.

BRENNAN
2000 Michael Brennan, *GAWK: Effective AWK Programming. A User's Guide for GNU Awk*, 2nd edition, Free Software Foundation Inc., 2000; disponibile online alla pagina: `http://www.gnu.org/software/gawk/manual/gawk.html`.

BURNARD
2005 Lou Burnard, *Metadata for Corpus Work,* in WYNNE 2005, pp. 30-46.

BUSA
1951 Roberto Busa SJ, *S. Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Milano, Bocca, 1951.

BUZZETTI
1999 Dino Buzzetti, *Rappresentazione digitale e modello del testo*, in *Il ruolo del modello nella scienza e nel sapere. Roma, 27-28 ottobre 1998*, Roma, Accademia Nazionale dei Lincei, 1999 "Contributi del Centro Linceo Interdisciplinare 'Beniamino Segre,'" 100, pp. 127-161.

CASAVECCHIA
2005 Sara Casavecchia, *Progettazione ed implementazione di corpora di lingua inglese basati sui newsgroup,* Università di Torino, Facoltà di Lingue, Tesi di Laurea 2004-2005.

ČERMÁK
1995 František Čermák, *Jazykový korpus: Prostřekek a zdroj poznání* [*Language Corpus: Means and Source of Knowledge*], in "Slovo a slovesnost" LVI (1995), pp. 119-140.
1997 František Čermák, *Czech National Corpus: a Case in Many Context*, in "International Journal of Corpus Linguistics" II (1997)[2], pp. 181-197.
2002 František Čermák, *Today's Corpus Linguistics. Some Open Questions*, in "International Journal of Corpus Linguistics" VII (2002), pp. 265-282.

CHRIST - SCHULZE
1996 Oliver Christ - Bruno Maximilian Schulze, *CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora*, in *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, herausgegeben von Helmut Feldweg und Erhard W. Hinrichs, Tübingen, Max Niemeyer Verlag, 1996 "Lexicographica. Series maior" 73; `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz`.

CHURCH - MERCER
1993 Kenneth W. Church - Robert L. Mercer, *Introduction to the special issue on computational linguistics using large corpora,* in "Computational Linguistics" XIX (1993)[1], pp. 1–24.

CIC
2006 *What is a Corpus?*, pagina web del *Cambridge International Corpus* (CIC), 2006, `http://www.cambridge.org/elt/corpus/what_is_a_cor pus.htm`.

CIURCINA - RICOLFI
¶ 7 Marco Ciurcina - Marco Ricolfi, *Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora*, in questo volume, pp. 127-132.

Commens Dictionary → BERGMAN - PAAVOLA 2003

CONNOR - UPTON
2004 *Applied Corpus Linguistics - A multidimensional Perspective*, edited by Ulla Connor and Thomas A. Upton, Amsterdam - New York, Editions Rodopi B.V., 2004 "Language and Computers".

CONTE
¶ 22 Amedeo Conte, *Valori normativi di verbi deontici in testi normativi*, in questo volume, pp. 363-370.

COOK - SEIDLHOFER
1995 *Principle & Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson,* edited by Guy Cook and Barbara Seidlhofer, Oxford - New York, Oxford University Press, 1995.

CORINO - MARELLO
2007 *i.s.* *Italiano di apprendenti. I Corpora VALICO e VINCA*, a cura di Carla Marello ed Elisa Corino, Perugia, Guerra, 2007, in corso di stampa.

CORINO - MARELLO - ONESTI
2006 *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 | Proceedings of the XII EURALEX International Congress. Torino, Italia, 6th-9th September 2006*, a cura di Elisa Corino, Carla Marello e Cristina Onesti, 2 voll., Alessandria, Edizioni dell'Orso, 2006.

CORPUSTYP → SINCLAIR 1996

CORRÉARD

2002 *Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins*, edited by Marie-Hélène Corréard, s.l. (U.K.), EURALEX, 2002.

CRESTI

2000 Emanuela Cresti, *Corpus di italiano Parlato*. Vol I. *Introduzione*, Vol II. *Campioni*, CD-ROM multimediale, Firenze, Accademia della Crusca, 2000.

CRYSTAL

1997 David Crystal, *The Cambridge Encyclopedia of Language*, second edition, Cambridge (UK), Cambridge University Press, 1997_2 [2002^r, 1987_1].

DAMASCELLI - MARTELLI

2002 Adriana Teresa Damascelli, Aurelia Martelli, *Corpus Linguistics and Computational Linguistics: an Overview with Special Reference to English*, Torino, Celid, 2002.

DE HAAN

1992 Pietre de Haan, *The optimum corpus sample size?*, in LEITNER 1992, pp. 3-19.

DEVOTO - OLI

2004 Giacomo Devoto - Gian Carlo Oli, *Dizionario della lingua italiana*, edizione 2004-2005 con CD-Rom, a cura di Luca Serianni e Maurizio Trifone, Firenze, Le Monnier, 2004.

DISC → SABATINI - COLETTI 2003.

DLP→ ALMEIDA COSTA - SAMPAIO E MELO 1999.

DOLI → DEVOTO - OLI 2004.

DRAE → REAL ACADEMIA 2004

DUBISZ

2004 *Uniwersalny słownik języka polskiego*, redakcja Stanisław Dubisz, Warszawa, Naukowe PWN, voll. 1-6 e CD ROM, 2004.

DWDS → ENZENSBERGER et alii 2003

EAGLES CORPUSTYP → SINCLAIR 1996

ENGWALL

1994 Gunnel Engwall, *Not Chance but Choice: Criteria in Corpus Construction*, in ATKINS - ZAMPOLLI 1994, pp. 49–82.

ENZENSBERGER et alii

2003 *DWDS: Das digitale Wörterbuch der deutschen Sprache des 20. Jh*, Kuratorium Hans Magnus Enzensberger, Wolfgang Frühwald, Gottfried Honnefelder, Wolf Lepenies, Christian Meier, Johannes Rau, Richard von Weizsäcker, Dieter E. Zimmer, Berlin-Brandeburgische Akademie des Wissenschaften, 2003, online alla pagina `http://www.dwds.de/`.

ΕΘΕΓ

2006 *Homepage* dell'*Εθνικός Θησαυρός Ελληνικής Γλώσσας (ΕΘΕΓ)* (Hellenic National Corpus (HNC)), 2006, `http://hnc.ilsp.gr/default.asp`.

FICKET - GUIGÓ

1993 James W. Fickett - Roderic Guigó, *Estimation of Protein Coding Density in a Corpus of DNA Sequence Data*, in “Nucleic Acids Research” XXI (1993)[12] 2837-2844. Disponibile anche online alla pagina `http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=309664&blobtype=pdf`.

FILLMORE
1992 Charles J. Fillmore, *"Corpus linguistics" or "Computer-aided armchair linguistics"*, in SVARTVIK 1992, pp. 35-60.

FLORIN
2004 Marcu Florin, *Marele Dicționar de neologisme*, ediția a VII-a revăzută, augmentată şi actualizată + CD ROM, Edituri Saeculum I. O., Bucureşti, 2004.

FONTENELLE
2004 Thierry Fontenelle, *Review* of Mitkov, 2003, in "International Journal of Lexicography" XVII (2004)[4], pp. 468-470.

FRANCIS
1964 W[inthrop] N[elson] Francis, *A standard sample of present-day English for use with digital computers. Report to the US Office on Education on Co-operative Research Project no E-007*, Providence, Brown University, 1964.
1982 W[inthrop] Nelson Francis, *Problems of Assembling and Computerizing Large Corpora*, in Johansson 1982, pp. 7-24.
1992 W[inthrop] N[elson] Francis, *Language Corpora B.C.*, in SVARTVIK 1992, in SVARTVIK 1992, pp. 17-32..

FRIES
1952 Charles Carpenter Fries, *The Structure of English; an Introduction to the Construction of English Sentences*, New York, Harcourt & Brace, 1952**.**

GARSIDE – LEECH – MCENERY
1997 *Corpus Annotation. Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech and Anthony McEnery, New York, Longman, 1997.

GDLI → BATTAGLIA [- BÀRBERI SQUAROTTI] 1961-2004.

GHADESSY - HENRY - ROSEBERRY
2002 *Small Corpus Studies and ELT: Theory and Practice*, edited by Mohsen Ghadessy, Alex Henry and Robert L. Roseberry, Amsterdam - Philadelphia, John Benjamins Pub Co, 2002 "Studies in Corpus Linguistics".

GLÜCK
2000 *Metzler Lexicon Spache*, zweite überarbeitet und erweiterte Auflage, herausgegeben von Helmut Glück, Stuttgart - Weimae, Verlag J. B. Metzler, 2000_2 [1993_1].

GRANGER
2004 Sylviane Granger, *Computer Learner Corpus Research*, in CONNOR - UPTON 2004, pp.121-145.

GRANGER - HUNG - PETCH-TYSON
2002 *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, edited by Sylviane Granger - Joseph Hung - Stephanie Petch-Tyson, Amsterdam - Philadelphia, John Benjamins Publishing Company, 2002 "Language Learning & Language Teaching Studies".

GREFENSTETTE
1999 Gregory Grefenstette, *Tokenization*, in VAN HALTEREN 1999, pp. 117-133.
1999a Gregory Grefenstette, *The World Wide Web as a resource for example-based machine translation tasks*, in *[ASLIB 21.] Proceedings of the 21st International Conference on Translating and the Computer*, London, November 1999, London, ASLIB Publications, 1999.

2002 Gregory Grefenstette, *The WWW as a Resource for Lexicography*, in CORRÉARD 2002, pp. 199-215.

GREFENSTETTE - KILGARRIFF → KILGARRIFF - GREFENSTETTE.

GREFENSTETTE - TAPANAINEN
1994 Gregory Grefenstette - Pasi Tapanainen, *What is a Word, What is a Sentence? Problems of Tokenization*, in the *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX '94)*, Budapest, Research Institute for Linguistics - Hungarian Academy of Sciences, 1994, pp. 79-87, disponibile online alla pagina `http://www.ling.helsinki.fi/~ta panain/tekeleet.html`.

HARRIS
1951 Zellig Harris, *Methods in Structural Linguistics*, Chicago, University of Chicago Press, 1951.

HEID
1998 Ulrich Heid, *Annotazione morfosintattica di corpora ed estrazione di informazioni linguistiche*, relazione al convegno *Annotazione morfosintattica di corpora e costruzione di banche di dati linguistici. Torino, 26-XI-1998*, inedita.

HUNDT - NESSELHAUF - BIEWER
2006 *Corpus Linguistics and the Web*, edited by Marianne Hundt, Nadja Nesselhauf and Carolin Biewer, Amsterdam, Editions Rodopi B.V., 2006 "Language & Computers".

HUNSTON
2002 Susan Hunston, *Corpora in applied linguistics*, Cambridge, Cambridge University Press, 2002.

JACKSON
2003 MacDonald P. Jackson, *Defining Shakespeare:* Pericles *as Test Case*, Oxford, Oxford University Press, 2003.

JOHANSSON
1982 *Computer Corpora in English Language Research. Proceedings of the 3rd ICAME Congress, Bergen 1981*, edited by Stig Johansson, Bergen, Norwegian Computing Center for the Humanities, 1982.
1991 Stig Johansson, *Computer Corpora in English Language Research*, in JOHANSSON - STENSTRÖM, 1991, pp. 3-6.

JOHANSSON - STENSTRÖM
1991 Stig Johansson, *Computer Corpora in English Language Research*, in *English Computer Corpora. Selected Papers and Research Guide*, edited by Stig Johansson and Anna-Brita Stenström, Berlin - New York, Mouton de Gruyter, 1991 "Topics in English Linguistics" 3.

JOJIĆ
2002 *Hrvatski enciklopedijski rječnik*, Ljiljana Jojić glavna urednica, Zagreb, Novi Liber, 2002.

JONES - TSCHIRNER
2006 Randall L. Jones - Erwin Tschirner, *A Frequency Dictionary of German. Core Vocabulary for Learners*, London - New York, Routledge, 2006 "Routledge frequency dictionaries".

KEHOE - RENOUF
2002 Andrew Kehoe - Antoinette Renouf, *WebCorp: Applying the Web to Linguistics and Linguistics to the Web*, in *[WWW 2002] Proceedings of the Eleventh International World Wide Web Conference. Honolulu, Hawaii, USA 7-11 May 2002*, Honolulu, 2002, online alla pagina `http://www2002.org/CDROM/ poster/67/`.

KENKYUSHA
2004 研究社新英和中辞典 [*Nuovo dizionario Kenkyusha inglese - giapponese ed. media*], disponibile online alla pagina `http://www.excite.co.jp/dic tionary/`.

KENNEDY
1998 Graeme Kennedy, *An introduction to corpus linguistics*, London - New York, Longman, 1998.

KILGARRIFF
2001 Adam Kilgarriff, *Web as a Corpus*, in RAYSON et alii 2001, pp. 342-344; poi anche in SAMPSON - MCCARTHY 2004, pp. 471-473.
2001a Adam Kilgarriff, *Comparing Corpora,* in "International Journal of Corpus Linguistics" VI (2001)[1] 1-37, disponibile anche online alla pagina `http://www.kilgarriff.co.uk/publications.htm`.

KILGARRIFF - GREFENSTETTE
2003 Adam Kilgarriff - Gregory Grefenstette, *Introduction to the Special Issue on the Web as Corpus*, in "Computational Linguistics" XXIX (2003)[3] 333-347, disponibile anche online alla pagina `http://www.kilgarriff.co.uk/publications.htm`.

KNOWLES - WILLIAMS - TAYLOR
1996 *A Corpus of Formal British English Speech*, edited by by Gerald Knowles, Briony Williams and Lita Taylor, London, Longman, 1996.

KOLDE
2006 Gottfried Kolde, *Korpuslinguistik - Corpus linguistics - Les linguistiques de corpus: Unterschiedliche Stile der Einführung in diese Disziplin? Lektüreerfahrungen eines der Einführung in die Korpuslinguistik bedürftigen weil blutigen Anfängers*, in NÄF - DUFFNER 2006, disponibile online alla pagina `http://www.linguistik-online.de/28_06/kolde.html`.

KOPOTEV
2003 М[ихаил] В. Копотев, *Корпусная лингвистика в Финляндии (обзор ресурсов)*, in "Научно-техническая информация. Сер. 2: Информационные системы и процессы" VI (2003) 37-43; versione inglese *Corpus Linguistics in Finnland: a Resource Survey*, in "Automatic Documentation and Mathematical Linguistics Translations of Selected Articles from Nauchno-Tekhnicheskaia Informatsiia" XXXVII (2003)[3] 35-42; entrambe disponibili online alla pagina `http://www.helsinki.fi/~kopotev/julkaisut.shtml`.

KORZEN
¶ 12 Iørn Korzen, *Mr. Bean e la linguistica testuale comparativa*, in questo volume, pp. 209-224.

KRAUSE
2005 *Nový Akademický Slovník Cizích Slov*, kolektiv autorů pod vedením Jiřího Krause, Praha, Accademia, 2005.

KUČERA
2002 Karel Kučera, *The Czech National Corpus: Principles, Design and Results*, in “Literary and Linguistic Computing” XVII (2002)^2, 245-257.

LEECH
1991 Geoffrey Leech, *The state of the art in corpus linguistics,* in AIJMER - ALTENBERG 1991, pp. 8-29.
1992 Geoffrey Leech, *Corpora and theories of linguistic performance*, in SVARTVIK 1992, pp. 105-122.
1997 Geoffrey Leech, *Introducing corpus annotation*, in GARSIDE - LEECH - MCENERY 1997, pp. 1-18.
2005 Geoffrey Leech, *Adding Linguistic Annotation,* in WYNNE 2005, pp. 17-29.

LEITNER
1992 *New Directions in English Language Corpora: Methodology, Results, Software Developments*, edited by Gerhard Leitner, Berlin, Mouton de Gruyter 1992 “Topics in English Linguistics”.

LEMNITZER - LOBIN
2004 *Texttechnologie: Perspektiven und Anwendungen*, herausgegeben von Lothar Lemnitzer und Henning Lobin, Tübingen, Stauffenburg, 2004 “Stauffenburg Einführungen”.

LEMNITZER - ZINSMEISTER
2006 Lothar Lemnitzer - Heike Zinsmeister, *Korpuslinguistik: eine Einführung*, Tübingen, Gunter Narr Verlag, 2006 “Narr Studienbücher”.

LENZ
2000 Susanne Lenz, *Korpuslinguistik*, Unveränd. Nachdr. d. 1. Aufl., Tübingen, Narr, 2006 [2000_1] “Studienbibliographien Sprachwissenschaft” 32.

LEWANDOWSKA-TOMASZCZYK - MELIA
1997 [*PALC ’97.*] *Proceedings of the First International Conference on Practical Applications in Language Corpora. Łódź, Poland, 10 - 14 April 1997*, edited by Barbara Lewandowska-Tomaszczyk and Patrick James Melia, Łódź; Łódź University, 1997.

LEWANDOWSKA-TOMASZCZYK - OSBORNE - SCHULTE
2001 Barbara Lewandowska-Tomaszczyk - John Osborne - Frits Schulte, *Foreign Language Teaching and Information and Communication Technology*, Frankfurt am Main, Peter Lang, 2001 “Łódź Studies in Language” 3.

LOVE
2004 Harold Love, *The Riddle of Pericles*, in “Times Litterary Supplement” - (2004 Aug. 13)$_{5289}$ 8-9.

LÜDELING - EVERT - BARONI
2006 Anke Lüdeling - Stefan Evert - Marco Baroni, *Using Web data for linguistic purpose*, in HUNDT - NESSELHAUF - BIEWER 2006, pp. 7-24.

LÜDELING - KYTÖ
2007 i.s. *Corpus Linguistics. An International Handbook*, edited by Anke Lüdeling - Merja Kytö, Berlin, Mouton de Gruyter, previsto per il 2007 “Series: Handbücher zur Sprache und Kommunikationswissenschaft | Handbooks of Linguistics and Communication Science”.

LÜDELING - KYTÖ - MCENERY
[2006] Anke Lüdeling - Merja Kytö - Tony McEnery, *Handbook on Corpus Linguistics - Outline*, presentazione di LÜDELING - KYTÖ 2007 *i.s.*, disponibile online alla pagina http://www2.hu-berlin.de/korpling/projekte/hsk/

LIZ → STOPPELLI - PICCHI 2001.

MAGNO CALDOGNETTO - COSI
2002 Emanuela Magno Caldognetto - Piero Cosi, *LIAV - Lessico dell'Italiano AudioVisivo. Corpus lessicale audiovisivo per l'analisi, la sintesi ed il riconoscimento bimodali dell'italiano parlato*, in "La comunicazione" XIII (2002), pp. 115-119; disponibile online alla pagina http://www.iscom.gov.it/documenti/files/rivista/2002_115.pdf. [numero speciale: *Atti della conferenza TIPI: Tecnologie Informatiche nella Processazione della Lingua Italiana*; versione online: http://www.iscom.gov.it/ contenuti.asp?ID=140&sID=24&xsID=81].

MANNING - SCHÜTZE
1999 Christopher D. Manning - Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge (Massachusetts) - London (England), The MIT Press, 2000^3 [1999_1].

MARELLO
1996 Carla Marello, *Le parole dell'italiano. Lessico e dizionari,* Bologna, Zanichelli, 1996.

MARCUS - SANTORINI - MARCINKIEVICZ
1994 Mitchell P. Marcus - Beatrice Santorini - Mary Ann Marcinkievicz, *Building a Large Annotated Corpus of English: The Penn Treebank*, in ARMSTRONG 1994, pp. 273-290. Disponibile online dalla homepage del PennTreebank al link ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz.

MEURMAN-SOLIN
2001 Anneli Meurman-Solin, *Structured Text Corpora in the Study of Language Variation and Change*, in "Literary & Linguistic Computing" XVI $(2001)^{1\text{-}2}$ 5-27.

MCARTHUR - MCARTHUR
1992 *The Oxford Companion to the English Language*, edited by Tom McArthur and Roshan McArthur, Oxford, Oxford University Press, 1992.

MCENERY
2003 Tony McEnery, *Corpus Linguistics*, in MITKOV 2003, pp. 448-463.

MCENERY - GABRIELATOS
2006 Tony McEnery - Costas Gabrielatos, *English Corpus Linguistics*, in AARTS -MCMAHON 2006, pp. 33-71.

MCENERY - WILSON
2001 Tony McEnery - Andrew Wilson, *Corpus Linguistics. An Introduction*, Edinburgh, Edinburgh University Press, 2001_2 [1996_1, 2005^r] "Edinburgh Textbooks in Empirical Linguistics".
2007 Tony McEnery - Andrew Wilson, *Corpus linguistics. ICT4LT Module 3.4*, last updated 27 February 2007, pagine web di ICT4LT, online a http://www.ict4lt.org/en/en_mod3-4.htm. [anche versione italiana, online alla pagina http://www.ict4lt.org/it/it_mod3-4.htm, aggiornata maggio 2000].

MEYER
2002 Charles F. Meyer, *English Corpus Linguistics: An Introduction*, Cambridge, Cambridge University Press, 2002 “Studies in English Language”.

MEYER -NELSON
2006 Charles F. Meyer - Gerald Nelson, *Data Collection*, in AARTS - MCMAHON 2006, pp. 93-113.

MIKHEEV
2003 Andrei Mikheev, *Text Segmentation*, in MITKOV 2003, pp. 201-218.

MITKOV
2003 *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov, Oxford, Oxford University Press, 2003.
2003a [Ruslan Mitkov], *Glossary*, in MITKOV 2003, pp. 727-760.

MNSZ
2005 *Magyar Nemzeti Szövegtár*, homepage, 2005, `http://corpus.nytud.hu/mnsz/index_hun.html` (*Hungarian National Corpus*, `http://corpus.nytud.hu/mnsz/index_eng.html`)

MUKHERJEE
2002 Joybrato Mukherjee, *Korpuslinguistik und Englischunterricht: Eine Einführung*, Frankfurt am Main, Peter Lang, 2002 “Sprache im Kontext” 14.

NÄF - DUFFNER
2006 *Korpuslinguistik im Zeitalter der Textdatenbanken / Corpus linguistics in the era of text data banks*, Heftherausgeber / Editors Anton Näf und | and Rolf Duffner, Linguistik online 28,3/06, `http://www.linguistik-online.de/ 28_06/`.

NEUHAUS
1988 Joachim H. Neuhaus, *Designing Lexical Databases*, in “Lexicographica” IV (1988), pp. 60-69.
1989 Joachim H. Neuhaus, *The Shakespeare Dictionary Database*, in “ICAME Journal” XIII (1989), pp. 3-11.

NKRJA
2003-06 *[Национальный корпус русского языка.] Что такое Корпус?*, pagina web 2003-2006, `http://ruscorpora.ru/corpora-intro.html`.

OCEL → MCARTHUR - MCARTHUR 1992

OED → SOANES - STEVENSON 2005

OOSTDIJK
1991 Nelleke Oostdijk, *Corpus Linguistics and the Automatic Analysis of English*, Amsterdam - Atlanta, Rodopi, 1991.

PAJUSALU - HENNOSTE
2002 *Tähendusepüüdja. Pühendusteos professor Haldur Õimu 60. sünnipäevaks | Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, toimetanud | edited by Renate Pajusalu ja | and Tiit Hennoste, Tartu, Tartu Ülikooli Kirjastus, 2002 “Tartu Ülikooli üldkeeleteaduse õppetooli toimetised | Publications of the Department of General Linguistics” 3.

PEIRCE
1906/31-58 Charles Sanders Peirce, *Prolegomena to an Apology for Pragmaticism*, 1906, in *Collected Papers of Charles Sanders Peirce*, 8 volumes, vols. 1-6, eds. Charles Hartshorne and Paul Weiss, vols. 7-8, ed. Arthur W. Burks. Cambridge, Mass.: Harvard University Press, 1931-1958, vol. IV, p. 537.

1980 → BONFANTINI ET ALII 1980.

PETŐFI
1988/96 Petöfi[103] János S[ándor], *La lingua come mezzo di comunicazione scritta: il testo*, in PETŐFI - VITACOLONNA 1996, pp. 66-107 [Prima edizione: Urbino, 1988, Centro internazionale di semiotica e linguistica dell'Università di Urbino; poi anche in inglese in *An Encyclopedia of Language*, edited by N[eville] E. Collinge, London - New York, Routledge, 1990, ¶ 7 pp. 207-243].
2004 Petöfi János S[ándor], *Scrittura e interpretazione. Introduzione alla testologia semiotica dei testi verbali*, Roma, Carocci Editore, 2004 "Università" 613.

PETŐFI - VITACOLONNA
1996 *Sistemi segnici e loro uso nella comunicazione umana. 3. La testologia semiotica e la comunicazione multimediale*, a cura di János S[ándor] Petöfi - Luciano Vitacolonna, Macerata, Università di Macerata, 1996 "Dipartimento di filosofia e scienze umane. Quaderni di ricerca e didattica" 17.

POWELL
2006 Chris Powell, *SEMiSUSANNE Corpus: Documentation*, pagina web `http://www.grsampson.net/SemiSueDoc.html`, Jan. 2006. [This Web version of Christopher Powell's SEMiSUSANNE readme file was prepared by Geoffrey Sampson on 17 Jan 2006].

PUSZTAI
2003 *Magyar értelmező kéziszótár. 2. átdolgozott és bővített kiadás* [A concise dictionary of definitions of Hungarian. 2nd revised and enlarged edition], főszerkesztő Pusztai Ferenc, Budapest, Akadémiai Kiadó, 2003.

PR → ROBERT 1991.

QUINE
1987 Willard van Orman Quine, *Quiddities: an Intermittently Philosophical Dictionary*, Cambridge (Mass.), the Belknap Press of Harvard University Press, 1987.

QUIRK - SVARTVIK → SVARTVIK - QUIRK

RAYMOND et alii
1992 Darrell R. Raymond, Frank W. Tompa and Derick Wood, *Markup Reconsidered*, paper presented at the *First International Workshop on Principles of Document Processing*, Washington DC, October 22-23, 1992; disponibile online alla pagina: `http://db.uwaterloo.ca/~drraymon/papers/markup.ps`.

RAYSON et alii
2001 *Proceedings of the Corpus Linguistics 2001 Conference. Lancaster University 29 March - 2 April 2001*, edited by Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja, Lancaster, University Center for Computer Corpus Research on Language, 2001 "UCREL Technical Paper" 13.

[103] Sic: Petőfi János Sándor, evidentemente rassegnato acché la /ö/ lunga dell'ungherese venga bistrattata dagli editori italiani, per prevenire maggiori danni si firma ormai in italiano "János Petöfi".

REAL ACADEMIA

2004 Real Academia Española, *Diccionario de la lengua española*, Vigésima segunda edición, disponibile online alla pagina `http://buscon.rae.es/draeI/`.

RENOUF

1987 Antoinette Renouf, *Corpus development*, in SINCLAIR 1987, pp. 1-22.

ROSEN

1972 Charles Rosen, *The Classical Style: Haydn, Mozart, Beethoven*, New York - London, Norton, 1972_2 [1971_1], p. 30.

ROBERT

1991 *Le petit Robert* par Paul Robert. *Dictionnaire alphabétique et analogique de la langue française*, rédaction dirigée par A. Rey et J. Rey-Debove, nouvelle édition revue, corrigée et mise à jour, Paris, Le Robert, 1991.

ROSSINI FAVRETTI

2000 *Linguistica e informatica.* Corpora*, Multimedialità e percorsi di apprendimento*, a cura di Rema Rossini Favretti, Roma, Bulzoni Editore, 2000.

2000a *Progettazione e costruzione di un* corpus *di italiano scritto: CORIS/CODIS*, in ROSSINI FAVRETTI 2000, pp. 39-56.

SABATINI

2006 Francesco Sabatini, *La storia dell'italiano nella prospettiva della* corpus linguistics, in CORINO - MARELLO - ONESTI 2006, pp. 31-37.

¶ ij Francesco Sabatini, *Storia della lingua italiana e grandi corpora. Un capitolo di storia della linguistica*, in questo volume, pp. xiij-xvj.

SABATINI - COLETTI

2003 *Il Sabatini Coletti. Dizionario della lingua italiana*, diretto da Francesco Sabatini e Vittorio Coletti, Milano, Rizzoli Larousse, 2003, 1 vol. + CD-ROM.

SAMPSON

1979 Geoffrey Sampson, *Liberty and Language*, Oxford - New York - Toronto - Melbourne, Oxford University Press, 1979.

1995 Geoffrey Sampson, *English for the Computer. The SUSANNE Corpus and Analytic Scheme*, Oxford, Clarendon Press, 1995.

1996 Geoffrey Sampson, *From Central Embedding to Corpus Linguistics*, in THOMAS - SHORT 1996, pp. 14-25 [poi anche, in versione modificata col titolo *From Central Embedding to Empirical Linguistics*, in SAMPSON 2001, pp. 13-23].

1997 Geoffrey Sampson, *Educating Eve. The 'Language Instinct' Debate*, London - New York, Cassel, 1997 [1999[r]] "Open Linguistics".

2001 Geoffrey Sampson, *Empirical Linguistics*, London - New York, Continuum, 2001 "Open Linguistics".

2004 Geoffrey Sampson, *Introduction* to SAMPSON - MCCARTHY 2004, pp. 1-8.

2004a Geoffrey Sampson, [*Foreword*] to *Charles Carpenter Fries, from The Structure of English. 1952*, in SAMPSON - MCCARTHY 2004, p. 9.

2006 Geoffrey Sampson, *The SUSANNE Analytic Scheme*, pagina web, `http://www.grsampson.net/RSue.html`. [last changed 25 Nov 2006].

SAMPSON - MCCARTHY

2004 *Corpus Linguistics. Readings in a Widening Discipline*, edited by Geoffrey Sampson and Diana McCarthy, London - New York, Continuum, 2004.

SASAKI - WITT

2004 Felix Sasaki - Andreas Witt, *Linguistische Korpora*, in LEMNITZER - LOBIN 2004, pp. 195-216.

SCHAUPP
2006 Annette Schaupp, *Entwicklung und Anwendung eines Metadatenmodells für das italienische Lernerkorpus Valico mit Fokussierung auf den Lernerhintergrund*, Stuttgart, Institut für maschinelle Sprachverarbeitung, 2006; Diplomarbeit Nr. 51, Prüfer HD Dr. Ulrich Heid, Zweitprüfer Dr. Helmut Schmid.

SCHERER
2006 Carmen Scherer, *Korpuslinguistik*, Heidelberg, Carl Winter, 2006.

SINCLAIR
1987 *Looking up: an Account of the COBUILD Project il Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, edited by J[ohn] M[cHardy] Sinclair, London - Glasgow, Collins ELT, 1987.
1991 John [McHardy] Sinclair, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press, 1991.
1996 J[ohn McHardy] Sinclair, *Preliminary Recommendations on Corpus Typology*, EAGLES Document EAG-TCWG-CTYP/P, Version of May, 1996, disponibile online alla pagina `http://www.ilc.cnr.it/EAGLES96/browse.html` (`corpustyp.ps` oppure `corpustyp.html`).
2000 John [McHardy] Sinclair, *Current Issues in Corpus Linguistics*, in ROSSINI FAVRETTI 2000, pp. 29-38.
2005 John [McHardy] Sinclair, *Corpus and Text. Basic Principles*, in WYNNE 2005, pp. 1-16.
2005a John [McHardy] Sinclair, *Appendix: How to Build a Corpus*, in WYNNE 2005 online alla pagina `http://www.ahds.ac.uk/creating/guides/linguistic-corpora/appendix.htm`

SLEX
1999 *SLEX99 - Elektronický lexikón slovenského jazyka*, s.l., Forma, 1999, online alla pagina `http://www.slex.sk/`.

SOANES - STEVENSON
2005 *Oxford Dictionary of English. Second Edition, Revisited*, edited by Catherine Soanes and Angus Stevenson, Oxford, Oxford University Press, 2005 [previous edition by Judy Pearsall and Patrick Hanks, *ibidem*, 1998].

SPERBERG-MCQUEEN - BURNARD
1999 *Guidelines for Electronic Text Encoding and Interchange. TEI P3 Text Encoding Initiative*. Revised reprint, edited by C. M[ichael] Sperberg-McQueen and Lou Burnard, Oxford May 1999. Prima edizione: *Guidelines for Text Encoding and Interchange*, edited by C. M. Sperberg-McQueen and Lou Burnard, Chicago and Oxford, Text Encoding Initiative, 1994. Poi versione P4: *The TEI Consortium: guidelines for electronic text encoding and interchange*, edited by C. M. Sperberg-McQueen and Lou Burnard, Oxford, Humanities Computing Unit, University of Oxford, 2002, *XML-compatible edition*, prepared by Syd Bauman [et alii]. Disponibile online alla pagina `http://www.tei-c.org/P4X/`.

SPINA
2001 Stefania Spina, *Fare i conti con le parole. Introduzione alla linguistica dei corpora*, Perugia, Guerra, 2001

STOPPELLI - PICCHI
2001 *LIZ 4.0. Letteratura italiana Zanichelli. CD-ROM dei testi della letteratura italiana*, a cura di Pasquale Stoppelli ed Eugenio Picchi, Bologna, Zanichelli, 2001, quarta edizione per Windows.

SVARTVIK
1992 *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82. Stockholm, 4-8 August 1991*, edited by Jan Svartvik, Berlin, Mouton de Gruyter, 1992 “Trends in Linguistics. Studies and Monographs” 65.
1992a Jan Svartvik, *Corpus linguistics comes of age,* in SVARTVIK 1992, pp. 7-13.

SVARTVIK - QUIRK
1980 *A Corpus of English conversation*, edited by Jan Svartvik and Randolph Quirk, Lund, C.W.K. Gleerup, 1980 “Lund studies in English” 56.

TFD
2007 *Corpus*, pagina online: `http://www.thefreedictionary.com/cor pus`, 2007 [last checked 15 February 2007].

TLFi → AA. VV. 2004a

THOMAS - SHORT
1996 *Using corpora for language research. Studies in the honour of Geoffrey Leech*, edited by Jenny Thomas and Nick Short, London - New York, Longman, 1996.

THOMPSON - HUNSTON
2005 *System and Corpus: Exploring Connections*, edited by Geoffrey Richard Thompson and Susan Hunston, London, Equinox Publishing, 2004 “Functional Linguistics”.

TOGNINI-BONELLI
2001 Elena Tognini-Bonelli, *Corpus Linguistics at Work*, Amsterdam - Philadelphia, John Benjamins Publishing Company, 2001 “Studies in Corpus Linguistics” 6.

TRIBBLE
1997 Chris Tribble, *Improvising Corpora for ELT: Quick and Dirty Ways of Developing Corpora for Language Teaching*, in LEWANDOWSKA-TOMASZCZYK - MELIA 1997. Disponibile online alla pagina `http://www.eisu.bham.ac.uk/johnstf/palc.htm`.

VAN HALTEREN
1999 *Syntactic Wordclass Tagging*, edited by Hans van Halteren, Dordrecht - Boston - London, Kluver Academic Publishers, 1999 “Text, Speech and Language Technology” 9.

VOLK
2001 Martin Volk, *Exploiting the WWW as a corpus to resolve PP attachment ambiguities*, in RAYSON et alii 2001, pp. 601-606, disponibile online alla pagina `http://www.ling.su.se/DaLi/volk/publications.html`.
2002 Martin Volk, *Using the Web as Corpus for Linguistic Research*, in PAJUSALU-HENNOSTE 2002. Disponibile online alla pagina `http://www.ling.su.se/DaLi/volk/publications.html`.

WAY - GOUGH
2003 Andy Way - Nano Gough, *wEBMT: Developing and Validating an Example-based Machine Translation System Using the World Wide Web*, in “Computational Linguistics” XXIX (2003)[3] 421-457, disponibile online alla pagina `http://www.computing.dcu.ie/~away/pubs.html`.

WIKIPEDIA
2006ja コーパス, pagina online: `http://ja.wikipedia.org/wiki/%E3%82%B3%E3%83%BC%E3%83%91%E3%82%B9`. [最終更新 2006年12月9日 (土) 09:05。].

2006pl *Korpus (językoznawstwo)*, pagina online: http://pl.wikipedia.org/wiki/Korpus_%28j%C4%99zykoznawstwo%29. [Tę stronę ostatnio zmodyfikowano 20:57, 19 wrz 2006].

2007a *Wikipedia: About*, pagina online: http://en.wikipedia.org/wiki/Wikipedia:About. [last modified 23:17, 14 February 2007].

2007cs *Jazykový korpus*, pagina online: http://cs.wikipedia.org/wiki/Korpus. [Stránka byla naposledy editována v 22:32, 24. 1. 2007].

2007de *Textkorpus*, pagina online: http://de.wikipedia.org/wiki/Textkorpus. [zuletzt am 7. Februar 2007 um 14:46 Uhr geändert].

2007en *Text corpus*, pagina online: http://en.wikipedia.org/wiki/Text_corpus. [last modified 07:58, 5 February 2007].

2007es *Corpus lingüístico*, pagina online: http://es.wikipedia.org/wiki/Corpus_ling%C3%BC%C3%ADstico. [modificada por última vez el 18:43, 7 feb 2007].

2007ru *Корпусная лингвистика*, pagina online: http://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%BA%D0%BE%D1%80%D0%BF%D1%83%D1%81. [Последнее изменение этой страницы: 22:32, 2 февраля 2007].

2007sk *Korpus (jazykoveda)*, pagina online: http://sk.wikipedia.org/wiki/Korpus_%28jazykoveda%29. [Čas poslednej úpravy tejto stránky je 17:47, 7. február 2007].

WYNNE
2005 *Developing Linguistic Corpora: a Guide to Good Practice*, edited by Martin Wynne, Oxford, Oxbow Books, 2005. Disponibile online alla pagina http://ahds.ac.uk/linguistic-corpora/.

ZANNI
¶ 6 Samantha Zanni, *Corpora elettronici e copyright. Lo stato legale della questione*, in questo volume, pp. 119-126.

CORPORA, NON-CORPORA (!), SOFTWARE E SITI DI RIFERIMENTO.

ADAM	http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=871/vers=ita
Answers.com	http://www.answers.com/
Athel.com	http://www.athel.com/corpus.html
Athenaeum Corpus	http://www.bmanuel.org/projects/at-HOME.html
BNC	http://www.natcorp.ox.ac.uk/ http://sara.natcorp.ox.ac.uk/lookup.html
Brown Corpus	http://en.wikipedia.org/wiki/Brown_Corpus http://ota.ahds.ac.uk/ (*search*)
Calgary Corpus	ftp://ftp.cpsc.ucalgary.ca/pub/projects/text.compression.corpus
CIC	http://www.cambridge.org/elt/corpus/international_corpus.htm

Canterbury Corpus	`http://corpus.canterbury.ac.nz/`
ČNK	`http://ucnk.ff.cuni.cz/` cfr. ČERMÁK 1997.
CodonCode Aligner	`http://www.codoncode.com/aligner/index.htm`
Corpora List	`http://listserv.linguistlist.org/archives/corpora.html`
Corpus Taurinense	`http://www.bmanuel.org/projects/ct-HOME.html`
CT → Corpus Taurinense	
CWB (& CQP)	`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`
EAGLES	`http://www.ilc.cnr.it/EAGLES96/home.html`
EKK	`http://www.cl.ut.ee/korpused/baaskorpus/index.php?lang=et`
Freiburg VKzAph → Lemnitzer - Zinsmeister 2006, pp. 124-125.	
ΕΘΕΓ	`http://hnc.ilsp.gr/en/`
HNK	`http://www.hnk.ffzg.hr/`
ICE	`http://www.ucl.ac.uk/english-usage/ice/`
ICT4LT	`http://www.ict4lt.org/`
IMS Stuttgart	`http://www.ims.uni-stuttgart.de`
Index Thomisticus	`http://www.corpusthomisticum.org/it/index.age`
Jus Jurium	`http://www.bmanuel.org/projects/ju-HOME.html`
LABLITA	`http://lablita.dit.unifi.it`
LABLITA - Campioni → CRESTI 2000.	
LCCPW	`http://bowland-files.lancs.ac.uk/lever/`
Linguistik Online	`http://www.linguistik-online.de/index.html`
LIAV → MAGNO CALDOGNETTO - COSI 2002	
LION	`http://collections.chadwyck.com/marketing/index.jsp`
LIZ 4.0 → STOPPELLI - PICCHI 2001	
LOB Corpus	`http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html#lob` `http://ota.ahds.ac.uk/` (*search*)
LLC	`http://ota.ahds.ac.uk/` (*search*) `ftp://ftp.cogsci.ed.ac.uk/pub/corpus-LLC/` → SVARTVIK - QUIRK 1980
LSE	`http://lse.umiacs.umd.edu:8080/`
METER Corpus	`http://www.dcs.shef.ac.uk/nlp/meter/index. html`
MNSz	`http://corpus.nytud.hu/mnsz/`

NKRJa	`http://ruscorpora.ru/index.html`
NUNC	`http://www.bmanuel.org/projects/ng-HOME.html`
OED	`www.oed.com`
OVI db testuale	`http://ovisun198.ovi.cnr.it/italnet/OVI/`
PennTreebank	`http://www.cis.upenn.edu/~treebank/`
Progetto Manuzio	`http://www.liberliber.it/biblioteca/index.htm`
Project Gutenberg	`http://www.gutenberg.org/wiki/Main_Page`
Semanticsarchive	`http://semanticsarchive.net/`
SEMiSUSANNE	`http://www.grsampson.net/Resources.html` cfr. POWELL 2006.
SGML	`http://xml.coverpages.org/sgml.html`
SNK	`http://korpus.juls.savba.sk/`
Susanne	`http://www.grsampson.net/Resources.html`
TACT	`http://www.chass.utoronto.ca/tact/`
TEI	`http://www.tei-c.org/`
TheFreeDictionary	`http://www.thefreedictionary.com`
Tottel's Misc. Corpus	`http://ota.ahds.ac.uk/` (*search*)
ToXgene	`http://www.cs.toronto.edu/tox/toxgene/index. html`
VALICO	`http://www.bmanuel.org/projects/br-HOME.html`
WaCky	`http://wacky.sslmit.unibo.it/doku.php`
WebCorp	`http://www.webcorp.org.uk/`

Wikipedia	`http://cs.wikipedia.org/wiki/`
	`http://de.wikipedia.org/wiki/`
	`http://en.wikipedia.org/wiki/`
	`http://es.wikipedia.org/wiki/`
	`http://ja.wikipedia.org/wiki/`
	`http://pl.wikipedia.org/wiki/`
	`http://ru.wikipedia.org/wiki/`
	`http://sk.wikipedia.org/wiki/`

Wordsmith's Tools	`http://www.lexically.net/wordsmith/`
X007 Benchmark	`http://www.comp.nus.edu.sg/~ebh/XOO7.html`
XML	`http://xml.coverpages.org/xml.html`

ULRICH HEID
(IMS Stuttgart)

4. Il CorpusWorkBench come strumento per la linguistica dei corpora[1]. *Principi ed applicazioni.*

0. INTRODUZIONE. L'uso di corpora è diventato in questi ultimi anni uno standard del lavoro descrittivo in linguistica, portando a quella disciplina spesso chiamata "linguistica dei corpora", "empirical linguistics", "corpus(-based) linguistics", ecc. I corpora sono considerati mezzi indispensabili per una descrizione più dettagliata delle lingue, per un'analisi quantitativa che permette di mettere in evidenza le preferenze distribuzionali, morfosintattiche e collocazionali delle parole e delle costruzioni linguistiche, e per il confronto di certi fenomeni tra tipi di testi, e tra differenti lingue, ovvero varietà, come nel lavoro dell'Università di Torino. L'approccio basato su corpora si sta attualmente evolvendo da una semplice metodologia ad una vera disciplina linguistica a sé stante (cfr. Lemnitzer - Zinsmeister 2005, che suggeriscono come la "linguistica dei corpora" abbia già acquisito questo stato).

Se i corpora sono così gli strumenti del linguista, la loro rappresentazione computazionale e la possibilità d'interrogarli in modo appropriato sono invece gli elementi indispensabili della tecnologia di base che tali strumenti sorregge. Siccome il progetto FIRB "L'italiano nella varietà dei testi" ha adottato il sistema CWB (Corpus WorkBench, cfr. Christ - Schulze 1996 e Christ et alii 1999) come rappresentazione dei suoi corpora e come motore di ricerca, presenteremo in quest'articolo il CWB e gli strumenti che contiene.

Non è nostra intenzione proporre un manuale dell'utente[2], né suggerire che il CWB sia l'unico sistema che permetta al linguista di lavorare seriamente con corpora. Piuttosto cercheremo di mettere in evidenza alcuni particolari caratteristiche della rappresentazione (cfr. § 1) e del motore di ricerca del CWB (cfr. § 2), e di discuterne alcuni aspetti d'uso sulla rete (cfr. § 3), partendo principalmente dalle esperienze del FIRB. In questa discussione faremo anche riferimento ad alcune interfacce già disponibili sulla rete, in primo luogo, ovviamente, a quella dei NUNC, sviluppata da Adriano Allora, di Torino.

1. CWB - UN SISTEMA PER LA LINGUISTICA DEI CORPORA. Il lavoro linguistico con corpora testuali, cioè con collezioni di dati linguistici sia scritti che orali, in genere è basato su un doppio uso di strumenti computazionali: da un lato, è necessaria una rappresentazione dei dati testuali e di tutto ciò che il linguista ha da dire su questi dati (cioè di tutti i tag ed il markup[3]), dall'altro è necessario un sistema che permetta all'utente l'identificazione di elementi specifici in questi dati, cioè un motore di ricerca.

1.1 CARATTERISTICHE GENERALI. L'aspetto della rappresentazione coinvolge due livelli, quello tecnico (la struttura informatica implementata) e quello logico (il modello linguistico computazionale sottostante). L'utente linguista in genere s'interessa meno della rappresentazione tecnica, purché essa metta a sua disposizione un accesso rapido, sicuro ed efficace ai corpora

[1] Nell'àmbito della giornata di studi, la comunicazione, di cui questo articolo è sostanziale rielaborazione, era sinteticamente intitolata *CQP: lo scheletro e il cuore tecnico di un corpus elettronico*.
[2] Manuali per utenti ed amministratori di corpora già esistono sulla rete: cfr. il CWB Users' Corner.
[3] Chiameremo qui *tagging* le "etichette" (introduzione di interpretazioni linguistiche) e *markup* le "annotazioni" (introduzione di metadata), giusta il sistema illustrato in questo volume da Barbera - Corino - Onesti ¶ 3.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 89-108.

ed al loro tagging e markup. Le varie alternative tecniche, ad esempio le banche dati, sistemi di indici su file testuali o su materiali codificati in XML, non saranno qui discusse in dettaglio. Bisogna però notare che il sistema CWB (Corpus WorkBench) discusso in questo articolo è stato concepito soprattutto per la lingua scritta, ragione per cui alcune sue specifiche tecniche possono anche costituire delle limitazioni riguardo al trattamento della lingua parlata. Ad esempio, rispetto ai sistemi basati su XML, non è possibile in CWB rappresentare strutture parzialmente sovrapposte (come due parlanti che parlano allo stesso tempo). In altri termini, è evidente che il modello tecnico e la scelta dei mezzi tecnici per la rappresentazione sono condizionati dall'uso linguistico previsto, dai tipi d'informazione da rappresentare e dal modello logico del corpus.

Il Corpus WorkBench è stato sviluppato nell'àmbito di progetti di linguistica dei corpora all'Università di Stoccarda. L'istituto di linguistica computazionale (IMS, Institut für maschinelle Sprachverarbeitung) iniziava la sua attività di lessicografia e di grammaticografia computazionali negli anni '90; ci si accorgeva allora assai rapidamente che l'uso di testi elettronici sarebbe stato necessario per verificare ipotesi grammaticali e lessicali. Tali tipi di verifiche sono possibili in primo luogo "in forma interattiva"; ma oltre a questa, come il linguista può interrogare il testo con una serie di richieste ed analizzarne dopo i loro risultati, così anche un programma dovrebbe poter applicare questa serie di richieste in modo automatico. Questo tipo di procedura è spesso chiamata "estrazione automatica di dati linguistici da corpora". Il CWB è concepito in modo da permettere ambedue questi tipi d'interrogazione, la "interattiva" e la "automatica"; in questa sede verrà discussa, però, solo la versione interattiva.

1.2 IL MODELLO CWB DI RAPPRESENTAZIONE DEL CORPUS. Analizzeremo ora il sistema CQP di rappresentazione con particolare riguardo all'aspetto sequenziale (cfr. § 1.2.1), all'introduzione di tagging e markup (cfr. § 1.2.2) ed alla "annotazione di regioni" (cfr. § 1.2.3).

1.2.1 ASPETTO SEQUENZIALE. Tradizionalmente[4], una frase può essere vista come una sequenza di parole, un paragrafo come una sequenza di frasi, un testo come una sequenza di paragrafi. Nello stesso modo, una frase parlata è ovviamente una sequenza di suoni linguistici, anzi una sequenza nel tempo.

	0
la	1
professoressa	2
spiega	3
la	4
parola	5
latina	6
"canis"	7
alla	8
studentessa	9

Tav 1: Testo sequenziale con numeri di posizione.

[4] Non si può, in questo articolo, fornire un riassunto della teoria e pratica della linguistica dei corpora, né della rappresentazione dei corpora. Presentazioni compatte di questi soggetti si trovano fra altro in Tognini - Bonelli 2001, Garside - Leech - McEnery 1997, McEnery - Xiao - Tono 2006. Una discussione approfondita di alcuni aspetti si trova in Barbera - Corino - Onesti ¶ 3, in questo volume.

Anche se, ovviamente, una tale rappresentazione puramente lineare è una semplificazione non priva di problemi (non è qui in discussione l'esistenza di elementi paradigmatici o comunque non lineari in un testo), essa può lo stesso utilmente servire come modello di base per una rappresentazione computazionale dei corpora. Infatti, una tale nozione di testo lineare viene usata in CWB come base della rappresentazione dei testi. Il sistema attribuisce numeri alle posizioni delle parole (o meglio, dei token[5]) nel corpus. Un esempio semplificato si trova nella tavola 1: l'inizio della frase riceve il numero "0", e ciascun token è numerato sequenzialmente.

Professoressa occupa dunque, nell'esempio della tavola 1, il posto che va dalla posizione 1 alla posizione 2, e *la*, presente due volte nella frase, occupa le posizioni da 0 a 1 e da 3 a 4. La rappresentazione sequenziale serve in particolare per il motore di ricerca: invece di esser costretto a cercare ogni elemento online nel testo (come avviene nelle ricerche effettuate con programmi come Perl), il CWB compila prima un indice, cioè una concordanza; quando poi il linguista cerca una parola, essa viene cercata non nel testo stesso, ma nell'indice, che ne contiene la rappresentazione numerica. È infatti più facile (ed efficace) computazionalmente cercare una forma nell'indice di quanto non lo sia cercarla, mettiamo, 300 volte in un testo. Il CWB utilizza dunque un indice posizionale precompilato (creato quando viene registrato il corpus per CWB). Ovviamente, c'è un prezzo da pagare per la relativa efficacia di questa tipo di ricerca: se vengono aggiunti nuovi testi ad un corpus esistente, bisogna ricompilare l'indice.

1.2.2 ETICHETTATURA ED ANNOTAZIONE. Nella parte precedente, il corpus è stato visto come un puro testo sequenziale, contentandosi di rappresentare segmentalmente il testo nel linguaggio di macchina. Però i linguisti sono interessati anche ad analizzare i testi e ad annotare i risultati di tale analisi. Nel caso più semplice, l'annotazione ("tag") è una etichetta ("label") aggiunta ad un elemento del testo.

È naturalmente possibile annotare testi a differenti livelli di astrazione descrittiva: al livello morfologico, al livello categoriale, sintattico oppure semantico. Un'etichettatura morfologica aggiunge a ciascuna forma flessiva una descrizione, per esempio in termini di genere, numero, ecc. Spesso etichette di questo tipo contengono anche informazioni categoriali. Questa etichettatura viene chiamata "etichettatura morfosintattica", oppure, nel gergo della *corpus linguistics*, "*part-of-speech tagging*" (POS-tagging). Un'etichettatura al livello semantico potrebbe invece aggiungere ad una parola il "synset" di WordNet corrispondente, ecc.

Non solo è possibile annotare elementi di un corpus a diversi livelli d'astrazione descrittiva, ma anche esiste una certa libertà per quanto riguarda la scelta di oggetti linguistici da annotare. Gli esempi discussi sopra (POS-tagging e tagging semantico) concernono sempre singole parole. Ma è altrettanto possibile annotare sequenze di parole (avverbi, preposizioni o congiunzioni plurilessematiche; locuzioni idiomatiche; ecc.), oppure gruppi sintagmatici (sintagmi nominali, aggettivali, preposizionali, ecc.), come accade nei "treebanks", cioè nei corpora annotati con alberi sintattici completi per ciascuna frase. Al di là della struttura frasale, ci possono anche essere annotazioni della struttura testuale (paragrafi, capitoli, titoli, ecc.). Nel modello del CWB, tutti questi tipi di sequenze annotate sono considerati come "regioni" ed annotati con attributi applicabili alla regione intera (cfr. *infra* § 1.2.3).

Tradizionalmente, i linguisti fanno una distinzione tra annotazioni linguistiche (tag) e metadata (markup): le annotazioni morfosintattiche, sintattiche, semantiche ecc. sono considerate come linguistiche, mentre i metadata danno informazioni sull'autore del testo, il contesto della produzione del testo, sul responsabile per la raccolta del testo nel corpus, e così via. Queste caratteristiche "esterne" possono essere utilissime nell'analisi sociolinguistica, nei corpora di apprendenti ecc. Alcuni preferiscono una triplice distinzione, tra annotazione linguistica, meta-

[5] Cfr. qui Barbera - Corino - Onesti ¶ 3, § 1.3.

data ed annotazione della struttura del testo; ed in genere non vi è molta chiarezza in queste distinzioni (cfr. qui Barbera - Corino - Onesti ¶ 3, § 1.4).

Dal punto di vista della rappresentazione del corpus, due principali modelli sono disponibili per tener conto del testo originale e delle annotazioni:

(j) ambedue possono essere notati sequenzialmente in un solo documento:

[1] `[<s>La/ART professoressa/NOM spiega/VER ... <s>]`

(ij) il testo e le annotazioni possono essere separati, in differenti documenti, con un sistema di rinvii (link, *pointers*, indici, ecc.) che esprime le relazioni tra gli oggetti di ciascun tipo (cfr. il metodo *stand-off* nei corpora rappresentati in XML).

In CWB è stato scelto il secondo modo di rappresentazione, più flessibile. Le posizioni numerate sono l'elemento di base della rappresentazione, e qualsiasi annotazione locale (cioè che si riferisca ad una sola posizione) è indicizzata su questa posizione. Dal punto di vista logico, questo è un modello a due dimensioni, come una tabella, in cui ogni posizione nel corpus può ricevere un numero variabile di tag[6].

	0	POS	GEND	NUM	LING	
la	1	DET	fem	sing	it	…
professoressa	2	NOM	fem	sing	it	…
spiega	3	VER		sing	it	…
la	4	DET	fem	sing	it	…
parola	5	NOM	fem	sing	it	…
latina	6	AGG	fem	sing	it	…
"canis"	7	UNK			lat	…
alla	8	P-DET	fem	sing	it	…
studentessa	9	NOM	fem	sing	it	…

Tav. 2: Testo sequenziale annotato

I tag sono interpretati come espressioni di attributi e valori, con l'attributo che definisce una dimensione di analisi linguistica (come la categoria, il numero, ecc.) ed i valori che indicano le istanze specifiche della dimensione (come nome, verbo, aggettivo, oppure singolare vs. plurale). Nella tavola 2 (*supra*) la medesima frase della tavola 1 viene ripresentata con le associazioni ad una annotazione di categoria (POS-tagging), a genere, a numero ed all'indicazione della lingua in cui la frase è espressa[7].

Anche le annotazioni sono indicizzate in CWB sulle posizioni, ragione per cui vengono chiamate "annotazioni posizionali" nel gergo del sistema. L'indice dunque contiene tutte le posizioni nel corpus, dove si trova, per esempio, un POS-tag `[pos = 'NOM']`. Siccome internamente vengono creati indici separati per ciascuna dimensione descrittiva (cioè per ogni "attributo"), è possibile interrogare il corpus per qualsiasi attributo, sia singolo che in combinazione.

---

[6] In linea di principio l'insieme dei tag applicabili è aperto, ma ci sono limitazioni pratiche.

[7] Si tratta solo di un *exemplum fictum*: le etichette categoriali (inventate) sono: DETerminante, NOMe, VERbo, AGGettivo, UNKnown (per la parola straniera) e P-DET, per indicare che si tratta della preposizione articolata.

1.2.3 ANNOTAZIONI DI REGIONI. Le annotazioni posizionali fanno riferimento solo a parole individuali (cioè, più precisamente, alle posizioni nel corpus dove si trovano i token in questione). Però, dal punto di vista linguistico, è necessario talvolta tenere conto anche di annotazioni per regioni: gruppi idiomatici (tipo *a meno che*, ecc.), gruppi sintagmatici (come gruppi aggettivali o nominali), frasi intere, post individuali all'interno di un thread di un newsgroup, materiali citati in un testo o quotati in un post di un newsgroup, ecc.

In CWB è stata introdotta allo scopo la possibilità di classificare le regioni e di indicarne l'inizio e la fine. Questo dispositivo (utilizzato dapprima solo per elementi della struttura testuale) è stato chiamato "annotazione strutturale". Riprendendo il solito campione, viene esemplificata nella tavola 3 l'annotazione della frase e del materiale citato:

	0	<s>	*inizio frase*
la	1		
…	..		
latina	6		
		<cit>	*inizio materiale citato*
"canis"	7		
		</cit>	*fine materiale citato*
alla	8		
studentessa	9		
	10	</s>	*fine frase*

Tav. 3: Testo sequenziale con annotazione strutturale della frase e del materiale citato.

Anche le annotazioni strutturali sono indicizzate, di modo che si possa chiedere al sistema di trovare tutte le regioni che contengano materiale citato, ecc.

Molto spesso, inoltre, il linguista s'interessa a fenomeni linguistici confinati all'interno di una frase: invece di essere costretto a fare riferimento alla punteggiatura, può così usare una richiesta vincolata "within sentence". Nello stesso modo è possibile, in linea di principio, cercare fenomeni transfrastici, specificando che la ricerca si applichi a due frasi, ecc.

1.3 LIMITAZIONI DEL MODELLO DI RAPPRESENTAZIONE. Il modello di rappresentazione CWB sopra accennato ha i suoi pregi e difetti. Vediamoli meglio.

1.3.1 SINTESI. In breve, il modello di rappresentazione del corpus di CWB ha le caratteristiche seguenti:

(j) è basato su una nozione di sequenzialità interna del corpus: il corpus è visto come una sequenza di singole posizioni;

(ij) permette l'annotazione di tutte le posizioni del corpus con coppie del tipo attributo/valore (se la dimensione descrittiva espressa da un attributo non è applicabile ad una posizione, il valore dell'attributo per questa posizione è "nil");

(iij) permette l'annotazione di regioni e la loro classificazione.

Un tale modello, molto semplice in linea di principio, si può implementare in modo efficiente; in CWB, il metodo principale è quello della creazione di indici separati per ciascuna classe di annotazione. Questi indici sono rappresentati internamente in modo compatto, usando il "coding Huffman" (cfr. Christ - Schulze 1996).

1.3.2 PROBLEMI E LIMITAZIONI. A livello teorico, il modello di CWB ha alcune implicazioni che potrebbero anche costituire delle limitazioni pratiche:

(j) È imperativa la presenza fisica del testo sulla stessa macchina dove opera il programma di ricerca. A differenza del sistema WebCorp, è impossibile con CWB lavorare su corpora distanti, virtuali. Per lavorare con testi presi dalla rete, è necessario prima scaricare questi testi, e poi trattarli in locale, come è stato fatto con i NUNC.

(ij) Il modello posizionale segue l'impostazione semplificatrice della linguistica computazionale nei confronti della nozione di parola: una parola è individuata come token in quanto sequenza ininterrotta di grafemi (cfr. *supra* Barbera - Corino - Onesti ¶ 3, § 1.3). Questo ovviamente crea difficoltà per le unità collocazionali (come *linguistica computazionale*), per le locuzioni polirematiche (come *anche se*, *senza che*), e per i nomi propri multilessicali (come *New York*, *Buenos Aires*). Tutti questi elementi devono essere trattati come "multiword".

Più problematico ancora è il caso delle forme fuse (*dammelo*, *farlo*) e delle preposizioni articolate (*alla*, *col*, *nelle*, ecc.): in teoria sarebbe necessario, per un trattamento adeguato, separarle, come viene fatto per l'italiano antico nel Corpus Taurinense[8], e per il catalano nel corpus dello IULA (Universidad Pompeu Fabra di Barcellona)[9]; ma non sempre è possibile o consigliabile farlo. Nei NUNC, ad esempio, non è per ora stato fatto.

Dato, inoltre, che non è più possibile modificare testo ed annotazioni una volta entrate nel sistema di concordanza (senza rifar tutto ogni volta daccapo), prima che un corpus venga caricato in CWB, il modello di tokenizzazione adotttato deve essere accuratamente studiato, ossia dovrebbe essere consistente, coerente e non-contradditorio.

(iij) L'uso delle annotazioni per le regioni è limitato dal fatto che non sono possibili né regioni ricorsive `[NP → N PRP NP]` né regioni sovrapposte (il caso già citato di due parlanti che parlino allo stesso tempo). Normalmente, il primo caso può essere evitato attraverso una modellizzazione iterativa `[NP → N PRP NP1]`, e il secondo si trova per fortuna raramente in materiali scritti.

1.3.3 IL LAVORO CON IL CWB. Per mettere a disposizione del linguista un corpus testuale tramite CWB sono necessarie alcune procedure preliminari.

In primo luogo, il testo deve essere fisicamente disponibile in locale e deve essere preparato con *tools* per la tokenizzazione (suddivisione in frasi e parole) e/o per l'annotazione morfosintattica (POS-tagging). Strumenti di pre-analisi di questo tipo sono disponibili in varie sedi; sulla homepage del TreeTagger ve ne sono per molte lingue fra cui l'italiano. Il risultato della loro applicazione al testo è la base a partire dalla quale sono creati gli indici CWB con un programma specifico, chiamato "encode" (cfr. il manuale nel CQP Users' Corner), che è una componente del sistema CWB.

2. IL MOTORE DI RICERCA CQP.

Il motore di ricerca CQP (Corpus Query Processor) è basato sulla rappresentazione logica (e sulla rappresentazione tecnica corrispondente) dei corpora in CWB, discussa sopra. Il motore di ricerca usa un "linguaggio regolare" (cioé basato su "espressioni regolari") per permettere di estrarre oggetti linguistici di varia natura. Tutti gli elementi indicizzati possono ovviamente essere usati nelle richieste ("query"), singoli o combinati.

---

[8] Dove tutti i "grafoclitici" (siano essi elementi pronominali od avverbiali nei gruppi verbali, ad es. *atartene* → *atar ÷te ÷ne*, od articoli nelle cosiddette preposizioni articolate, *della* → *de ÷lla*) sono tokenizzati separatamente; cfr. qui oltre Barbera ¶ 8, § 5.2.7 e nota 55.

[9] Così la forma *dels* (*de* + *els*) nella *frase amb una versiò simplificada de/de els/el codis d'error* ('con una versione semplificata dei codici d'errore') è scissa in due parti.

2.1 ELEMENTI DEL LINGUAGGIO DI RICERCA. La notazione del linguaggio d'interrogazione è ispirata alla notazione in attributi e valori, usata anche in linguistica; ciascun termine fa riferimento ad una posizione, e va concepito come una descrizione vincolata (*constraint*) di una posizione. Per esempio, la query in [2a] cerca tutte le forme taggate come verbi al presente (ne riportiamo alcuni risultati)[10].

[2a] `[pos = "VER:pres"`[11]`];`[12]
[2b] `ha, tengo, siamo, sai, è, ...`

È possibile, in questo linguaggio, servirsi della negazione di ogni valore (la negazione, ossia, è intesa come un coefficiente della lista di valori dichiarati); dunque "`[pos! = "VER:pres"]`" significa 'qualsiasi POS, salvo `"VER:pres"`' ed ha, ad esempio, `[pos = "NOM"]` oppure `[pos = "VER:impe"]` come suoi possibili risultati.

Siccome si tratta di un linguaggio regolare, anche più vincoli possono essere espressi, tramite espressioni congiuntive ("and") ed alternative ("or"). Un esempio si trova in [3a], con alcuni risultati estratti dai NUNC cucina, in [3b-d]:

[3a] `[pos= "VER:impe" & word = ".+ci|.+lo"];`
[3b] `Coprite il fondo di una teglia con sale grosso e` **`piantateci`** `i carciofi .` NUNC-IT Cooking,
[3c] `[...] ,` **`mettici`** `anche due cucchiai di pangrattato dentro ,` NUNC-IT Cooking,
[3d] `Se hai dei dubbi sul Grand Marnier` **`sostituiscilo`** `con il brandy .` NUNC-IT Cooking.

Ovviamente, le richieste possono anche applicarsi a sequenze di parole. Un esempio semplificato si trova in [4a]: la parola *senza*, seguita da un verbo all'infinito, un articolo, un aggettivo (facoltativo) ed un nome.

[4a] `[word = "senza"] [pos = "VER:infi"] [pos = "DET.*"] [pos = "ADJ"]? [pos = "NOM"];`
[4b] `posti buoni per mangiare buon pesce` **`senza spendere una fortuna`** NUNC-IT Cooking,
[4c] `il quale,` **`senza annusare il vino`** `ci ha cambiato immediatamente la bottiglia` NUNC-IT Cooking,
[4d] `reintrodurre i linfociti ... ,` **`senza interrompere le altre terapie`** NUNC-IT Cooking,
[4e] `pur` **`senza perdere le proprie caratteristiche`** `distintive` NUNC-IT Cooking.

Ci sono diversi dispositivi di sottospecificazione. Un sistema molto semplice sfrutta la struttura gerarchica del tagset (per la nozione cfr., in questo volume, Barbera ¶ 8, § 3 e sottoparagrafi), e come questa gerarchia si traduce nelle abbreviazioni delle etichette ("labels"). In tale senso, `"[pos = DET.*]"` in [4a] significa 'qualsiasi tipo di DET', perché i due sottotipi completamente specificati sono scritti `"DET:def"` e `"DET:indef"`. Un modo di lasciare una richiesta morfologicamente e sintatticamente sottospecificata è usare il simbolo della posizione arbitraria, "`[]`", che richiede che ci sia una parola, senza però specificarne la natura. Infine, anche le espressioni di alternative rivestono a volte un aspetto di sottospecificazione: la que-

---

[10] In tutti gli esempi italiani dati in questo paragrafo ci serviremo dei NUNC con il tagset attualmente impiegato; un elenco dei tag è presente sulla homepage dei NUNC.
[11] Nelle query di CQP può esere usato indifferentemente l'apice semplice "'" od il doppio """.
[12] Il "punto e virgola" è indispensabile alla fine di ogni comando di CQP; nell'interfaccia web, tuttavia, non è necessario introdurlo manualmente perché provvede a ciò automaticamente già il software.

ry in [5a] permette, tra la congiunzione *se* ed il verbo, sia pronomi che un gruppo nominale di articolo e sostantivo. Alcuni esempi di questi due tipi di "sintagmi nominali", estratti dai NUNC, sono dati in ([5b-d]).

[5a] `[word = "se"] ( ( [pos = "PRO:indef" | pos = "PRO:pers"] ) | ( [pos = "DET.*"] [pos = "NOM"] ) ) [pos = "VER.*"];`

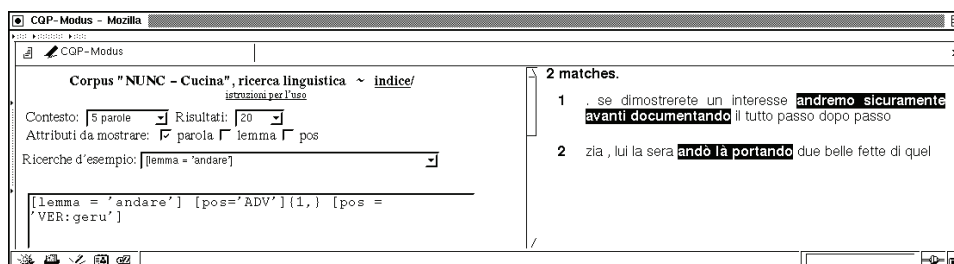
[5b] `volevo sapere` **`se mi sappiate`** `dire un buon ristorante` NUNC-IT Cooking,

[5c] `vorrei sapere` **`se altri hanno`** `avuto modo di` NUNC-IT Cooking,

[5d] **`se il Margaux fosse`** `venduto alla Coop per 8 Euro e forse quindi per tutti non sarebbe più Margaux` NUNC-IT Cooking.

Non è possibile qui dare una descrizione dettagliata e completa del linguaggio di interrogazione CQP; una tale descrizione si trova comunque nel manuale dell'utente. Dal punto di vista del linguista è utile notare che è possibile fare query generali o specifiche, attraverso "or" e/o valori sottospecificati; spesso, anzi, un fenomeno non si può subito circoscrivere in modo molto stretto, e dunque può essere utile iniziare una serie di esplorazioni nel corpus con query abbastanza generiche; l'analisi dei primi risultati può così permettere al linguista di riformulare la sua query in modo più specifico.

2.2 LA VISUALIZZAZIONE DEI RISULTATI. Essendo concepito in primo luogo per l'uso interattivo, il CQP contiene un'interfaccia semplice ed efficace per la visualizzazione dei risultati di ricerche in corpora. L'uso in rete anziché in locale comporta la perdita di alcune potenzialità di interattività e finezza di ricerca; tuttavia, uno dei grandi vantaggi dell'uso del CQP in rete è la possibilità di costruire interfacce specifiche, più adatte ad usi ed utenti specializzati (cfr. § 3).

Per quanto riguarda la visualizzazione dei risultati, bisogna distinguere due parti del testo nel quale viene fatta una ricerca: gli elementi che soddisfanno le condizioni della richiesta, ed il "contesto" che si trova attorno. Negli esempi [3b-e] soprariportati, solo la parte in grassetto corrisponde alla query ("*senza* seguito da un infinitivo e da un gruppo articolo + nome"), mentre il materiale che precede e segue è una parte (arbitrariamente selezionata dal presente autore) del "contesto". La distinzione tra "risposta" (quello che letteralmente corrisponde a quanto chiesto nella query) e "contesto" (il co-testo della risposta) è sistematicamente attuata dal CQP, che la rende graficamente nell'interfaccia standard (in locale) usando parentesi uncinate ("<...>") e colore invertito (bianco su nero) per la "risposta" vera e propria. Nell'interfaccia web dei NUNC, invece, tale "risposta" viene indicata con un evidenziazione colorata (cfr. Tav. 4).

Tav. 4. Interfaccia dei NUNC con distinzione tra risposta e contesto.

Ora il linguista potrebbe essere interessato nel contesto più ampio di tutta la frase complessa; potrebbe ad esempio volere l'esempio [4b] catturato come [4b'], oppure potrebbe desiderare vedere l'articolo del newsgroup per intiero, cfr. [4b] come [4b'']:

[4b'] Conosci altri posti buoni per mangiare buon pesce **senza spendere una fortuna** NUNC-IT Cooking,

[4b''] Re : Antica Osteria di Vico Palla (GE) sei genovese , vero ? Conosci altri posti buoni per mangiare buon pesce **senza spendere una fortuna** ? Preferisco la qualità all' apparenza .... 1 grazie , ciao NUNC-IT Cooking.

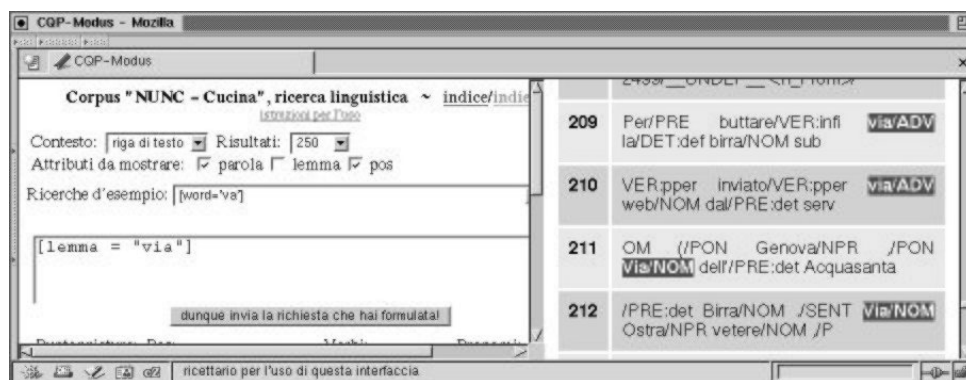
Il CQP permette difatti la definizione interattiva del contesto, soprattutto se il corpus interrogato ha marche di frase, unità testuale (post, paragrafo, ecc.). Per ottenere ciò, prima che di fare una richiesta è necessario impostare il contesto con il semplice comando

[6a] `set c 1 s` ‘set context 1 sentence’
[6b] `set c 5 s` ‘set context 5 sentences’, ecc.,

oppure ricorrendo al numero di grafemi (se non vi sono marche testuali disponibili)

[6c] `set c 20` ‘set context 20 graphemes’.

Per studi di linguistica testuale può essere, tra l’altro, importante accedere ad elementi della struttura testuale più ampi della frase (cfr. ad es. in questo vol. Cignetti ¶ 11).

Oltre alla visualizzazione (“`show`”[13]) in termini esclusivamente di “risultati-contesto”, come sopra esemplificato, è possibile formulare l’informazione visualizzata anche in termini di attributi e valori codificati. In un corpus annotato con valori di categoria (POS) e lemma (ossia lemmatizzato e POS-taggato), infatti, possono essere visualizzati oltre alle forme delle parole (token) anche le POS od i lemmi od ambedue, cfr. gli esempi tratti dei NUNC generici in [7], dove [b] visualizza (“`show`”) solo “`word`”, [c] anche “`lemma`”, [d] anche “`POS`”, ed [e] “`POS`” e “`lemma`” insieme, e cfr. la tav. 5.

Tav. 5. Interfaccia dei NUNC cucina con visualizzazione dei POS-tag.

[7a] `[lemma = "andare"] [word = "a"] [pos = "VER:infi"];`

[7b] kiavik viene trasferito dall ' azienda per la quale lavora , in marocco . e **va a vivere** lì . come lo vedranno i marocchini ? NUNC-IT Generic I,

[7c] Un/un po/<unknown> '/' più/più tardi/tardi ,/, un/un '/' altra/altro persona/persona lo/lo **va**/andare **a**/a **trovare**/trovare ./. NUNC-IT Generic I,

[13] In locale ciò si ottiene modificando le impostazioni di visualizzazione con il comando “`show`”; nell’interfaccia web cioè è reso possibile (de)cliccando delle caselle appositamente predisposte.

[7d] `Adesso/ADV .../PON gli/DET:def animali/NOM per/PRE cibarsi/VER:infi lo/DET:def fanno/VER:pres e/CON nessuno/ADJ gli/PRO:pers` **`va`**`/VER:pres` **`a`**`/PRE` **`dire`**`/VER:infi che/PRO:rela non/ADV è/VER:pres etico/ADJ )/PON` NUNC-IT Generic I,

[7e] `Per/PRE/per scoprirlo/VER:infi/scoprire ,/PON/, bisogna/VER:pres/bisognare` **`andarsi`**`/VER:infi/andare` **`a`**`/PRE/a` **`rileggere`**`/VER:infi/rileggere i/DET:def/il post/PRE/post precedenti/NOM/precedente ,/PON/, il/DET:def/il che/PRO:rela/che non/ADV/non sempre/ADV/sempre è/VER:pres/essere facile/ADJ/facile ,/PON/,` NUNC-IT Generic II.

Inoltre esistono diversi modi di ordinare i risultati[14], oltre a quello in cui compaiono nel corpus (cfr. [8b]): in ordine alfabetico, a seconda di qualsiasi attributo, ecc. Nello stesso modo si possono produrre anche liste di frequenza, sia per elemento singolo che per cooccorrenza di due elementi (cfr. [8]). Gli esempi [8-9], presi dal corpus di dibattiti del parlamento europeo (EUROPARL/MLCC - ELDA W0023), partono dalle richieste indicate e ne estraggono oggetti del verbo *risolvere*, oppure sequenze di verbo ed oggetto.

```
[8a]  [lemma = "risolvere"] [pos = "DET.*"] [pos = "NOM"];
[8b]  11475:    risolvendo le questioni                MLCC W0023,
      13803:    risolvere i problemi
      110454:   risolvere la controversia
[8a'] group Last matchend lemma;
[8b'] problema                | 63                     MLCC W0023,
      controversia            | 10
      questione               | 7
      conflitto               | 4
[9a]  [pos = "VER:infi"] [pos = "DET.*"] [pos = "NOM"];
[9a]  promuovere sviluppo     | 57                     MLCC W0023.
      risolvere problema      | 54
      raggiungere obiettivo   | 43
```

Siccome, poi, tutte le query CQP possono ricevere un “nome” (una variabile interna), è possibile anche fare riferimento a risultati precedenti, combinandoli oppure sottraendoli uno dell’altro (e questo permette anche la costruzione di query a cascata).

2.3 UN ESEMPIO DI QUERY LINGUISTICA: I GRUPPI VERBO+OGGETTO. Nel prosieguo saranno discusse alcune query CQP che permettono di estrarre coppie composte da un verbo e il sostantivo che è la testa del suo oggetto diretto. Queste tipo di coppie sono centrali nello studio delle collocazioni basato su corpora. Gli esempi provengono da un corpus di testi legislativi[15] reso disponibile all’autore dall’istituto per il multilinguismo e la comunicazione specialistica dell’Accademia Europea di Bolzano[16].

Nell’estrazione delle collocazioni, infatti, di solito si parte dalle coppie verbo + oggetto diretto, come ad es. *applicare* + *pena*, *mantenere* + *segreto professionale*, *prescrivere* + *azione*, *risarcire* + *danno* ecc. Essere in grado di valutare se una coppia sia stata trovata in forma attiva o passiva può, inoltre, essere utile in una fase ulteriore della ricerca. Nei testi analizzati, i candidati collocazionali appaiono sia all’attivo (*X applica [...] la/una ... pena*) che in costruzioni

[14] Le tecniche di “grouping”, in genere, sono però possibili solo in locale e non attraverso gli interfaccia web.
[15] Più accuratamente, si tratta del corpus non liberamente disponibile del progetto LexAlp, composto di 176 leggi e fonti normative del Friuli Venezia Giulia che riguardano la pianificazione territoriale e lo sviluppo sostenibile.
[16] L’autore desidera ringraziare A. Abel, I. Ties e V. Lyding per il loro supporto e per avergli messo a disposizione i corpora giuridici dell’EURAC.

passive con due ordini di parole differenti: nelle leggi, infatti, si trova sì l'ordine standard (*la pena è applicata*), ma anche l'ordine invertito (*viene applicata ... la pena..., se ...*)[17].

Per cercare i candidati collocazionali all'attivo, si può formulare l'ipotesi iniziale che qualsiasi forma attiva possa essere seguita da un gruppo nominale oggetto. Tra il verbo ed il sintagma nominale oggetto possono apparire avverbi e/o gruppi preposizionali (cfr. l'esempio [12c]). Siccome i corpora non sono annotati al livello di gruppi sintagmatici, è necessario formulare un modello approssimativo di un sintagma nominale[18] nei termini del materiale categorialmente etichettato, cfr. [10].

[10]
```
[pos = "DET.*"] [pos = "ADJ"]{0,3} [pos = "NOM"];
```

La sequenza di verbo, avverbi facoltativi e gruppo nominale (cfr. [11]) è completata in [12a] dal gruppo preposizionale facoltativo che può intervenire a sinistra del SN oggetto[19].

[11]
```
[pos = "VER.*" & pos != "VER:pper" & lemma != "essere"] [pos =
"ADV"]? ([pos = "DET.*"] [pos = "ADJ"]{0,3} [pos = "NOM"]);
```

[12a]
```
[pos = "VER.*" & pos != "VER:pper" & lemma != "essere"] [pos =
"ADV"]? ([pos = "PRE.*"] []? [pos = "ADJ"]{0,3} [pos = "NOM"]
[pos = "ADJ"]?)? ([pos = "DET.*"] [pos = "ADJ"]{0,3} [pos =
"NOM"]);
```
[20]

Le query [11] e [12a] restringono ulteriormente la forma verbale, che non può essere un participio passato né può appartenere al lemma *essere*. In [10] il gruppo nominale è modellizzato in modo semplicissimo, e certo insufficiente per casi più complessi, però [12a] permette già di estrarre casi relativamente complessi come quelli in [12b-d]:

[12b] Il venditore deve altresì **risarcire** al compratore i **danni** [...] LexAlp,

[12c] [...] prosegue lo scopo di **coordinare** in un contesto unitario le **azioni** promozionali di enti pubblici [...] LexAlp,

[12d] [l' articolo ...] **comporta** per i comuni inadempienti il **divieto** di rilasciare [...] LexAlp.

Purtroppo, allo stesso tempo vengono estratte anche frasi che contengono un predicato plurilessematico come *rimanere in vigore*, cfr. [12e-f]. Se non vengono identificati, questi casi producono "falsi positivi" come *rimanere in vigore* per *rimanere + divieto*. Però la sola possibilità per identificare predicati plurilessematici di questo tipo ci sembra l'uso di un dizionario che ne contenga i più importanti. In realtà, la collocazione tra *divieto* e la multiword *rimanere in vigore* sono elementi affatto correnti nel linguaggio idiomatico e collocazionale dell'amministrazione.

[12e] **Rimangono** in vigore gli ulteriori **divieti** stabiliti dall' articolo... LexAlp,

[12f] [...] **restano** in vigore i **divieti** di percorrenza ... LexAlp.

---

[17] Naturalmente, poi, le frasi da analizzare comportano non solo tempi verbali sintetici (presente, futuro, imperfetto, passato remoto), ma anche costruzioni con participi. Queste ultime non sono qui prese in considerazione.

[18] Questo modello vale anche per i gruppi preposizionali.

[19] Da notare che questo gruppo preposizionale permetterebbe anche un aggettivo postnominale.

[20] Negli interfaccia web (come ad esempio in quello dei NUNC) di solito non è possibile articolare la query su più righe (come un normale listato Perl od AWK). In locale, da terminale, ciò è invece possibile, con gran vantaggio della leggibilità e della compilazione di query complesse. La query in [12a] da terminale sarebbe pertanto:

[12a']
```
[pos = "VER.*" & pos != "VER:pper" & lemma != "essere"]
[pos = "ADV"]?
([pos = "PRE.*"] []? [pos = "ADJ"]{0,3} [pos = "NOM"] [pos = "ADJ"]?)?
([pos = "DET.*"] [pos = "ADJ"]{0,3} [pos = "NOM"]);
```

Per il passivo, vengono riutilizzati in parte gli stessi elementi della query per l'attivo, [10]. La parte verbale può essere modellizzata in maniera sottospecificata (cfr. [13]), e per il gruppo nominale diventato ora il soggetto del verbo passivo sono previsti facoltativi gruppi preposizionali postnominali, cfr. la modellizzazione dell'ordine invertito delle parole, in [14]:

[13]
```
[lemma = "essere|venire"] [word = "stat[o|a|i|e]"] [pos =
"VER:pper"];
```

[14].
```
[lemma = "essere|venire"] [word = "stat[o|a|i|e]"] [pos =
"VER:pper"] ([pos = "PRE.*"] [pos = "DET.*"]? [pos = "ADJ"]{0,3}
[pos = "NOM"])? [pos = "DET.*"] [pos = "ADJ"]{0,3} [pos = "NOM"]
([pos = "PRE.*"] [pos = "DET.*"]? [pos = "ADJ"]{0,3} [pos =
"NOM"]){0,2};
```
[21]

Applicate al materiale giuridico menzionato, queste procedure dànno candidati collocazionali come: *risarcire* + *danno*, *indennizzare* + *danno*, *attuare* + *azione*, *esercitare* + *azione*, *applicare* + *pena*, *contenere* + *divieto* (estratti da frasi attive); *concedere* + *contributo*, *svolgere* + *funzione*, *presentare* + *domanda*, *adottare* + *deliberazione* (estratti da frasi passive); ecc.

3. CQP IN RETE. Come accennato, il CQP può essere usato in modo interattivo ed in modo automatico. Entrambi i modi sono disponibili su un'architettura client/server, nella quale il server contiene i corpora ed il motore di ricerca, mentre il client contiene un'interfaccia utente.

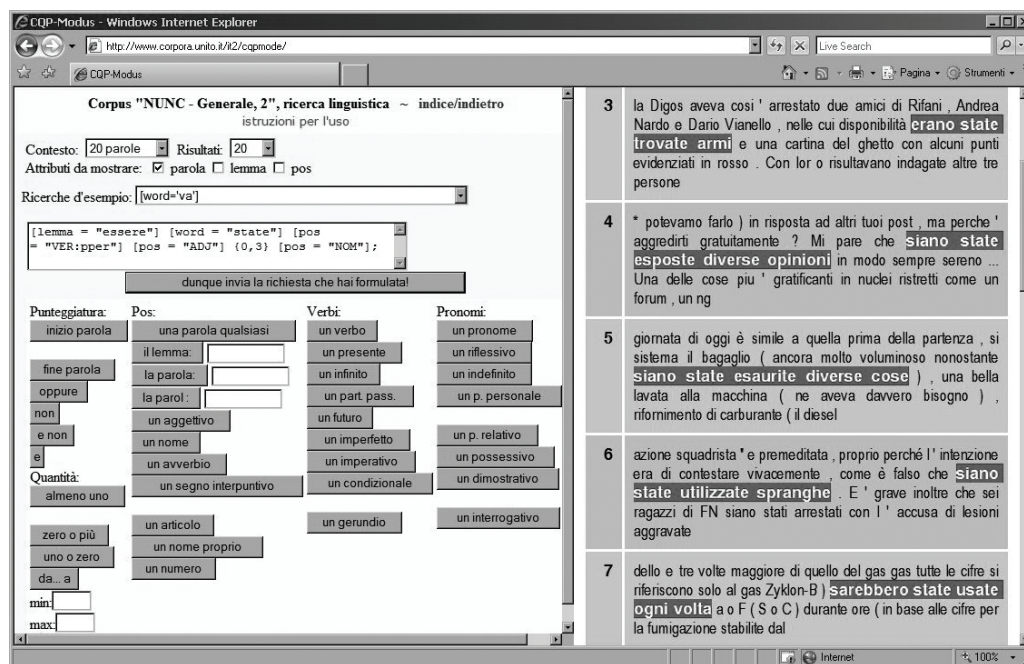
Per l'uso interattivo in rete, ci sono diversi aspetti che meritano una discussione più ampia. Nei paragrafi seguenti ci proponiamo di discuterne alcuni che concernono soprattutto i metodi per aiutare l'utente a formulare query, la messa a disposizione di differenti corpora, e la visualizzazione dei risultati. Siccome il CWB è stato usato, negli ultimi anni, in istituzioni di differenti paesi, potremo fare riferimento a diverse realizzazioni.

Il CWB è stato, per esempio, usato nella Linguateca portoghese, per l'esplorazione del corpus AC/DC di 200 milioni di parole. Per il danese, i corpora Korpus 90 e Korpus 2000 sono liberamente consultabili, così come lo sono i NUNC per l'italiano, il francese, lo spagnolo ed il tedesco. Il CWB supporta anche il corpus catalano IULA di Barcellona, un corpus misto che contiene sia dati provenienti dalla lingua generica che testi di diverse LSP. Infine, il corpus bosniaco (KBTUO: *Korpus bosanskih tekstova na Univerzitetu u Oslu*)[22] ed i corpora paralleli OPUS dell'Università di Oslo, il corpus svedese del progetto PAROLE, ed altri ancora., sono ugualmente basati su CWB.

3.1 INTERFACCE PER DIFFERENTI TIPI D'UTENTI. L'uso di corpora è diventato ormai una pratica abbastanza consolidata in linguistica; però non tutti i linguisti utilizzano un sistema d'interrogazione regolarmente; e quindi, se l'uso di un sistema messo a disposizione sulla rete apparisse troppo complesso, l'utente occasionale tenderà probabilmente ad evitarlo *tout court*.

[21] Che da terminale si leggerebbe:

[14']
```
[lemma = "essere|venire"] [word = "stat[o|a|i|e]"] [pos = "VER:pper"]
([pos = "PRE.*"] [pos = "DET.*"]? [pos = "ADJ"]{0,3} [pos = "NOM"])?
[pos = "DET.*"] [pos = "ADJ"]{0,3} [pos = "NOM"]
([pos = "PRE.*"] [pos = "DET.*"]? [pos = "ADJ"]{0,3} [pos = "NOM"]){0,2};
```

[22] Questo corpus può però essere usato solo con verifica d'accesso via password.

Tav. 6. Interfaccia con modulo di autocomposizione dei NUNC[23].

Bisognerebbe dunque concepire anche interfacce semplici e facili da usare per chi non ne facesse uso abituale. Idealmente, sarebbe necessaria una serie d'interfacce alternative per lo stesso sistema di ricerca / corpus, almeno una semplificata per utenti occasionali ed una avanzata per "specialisti"[24]. Per questo, ad esempio, l'interfaccia in rete del tedesco Bundestag Corpus dell'IMS si presenta in due modi, uno assai semplice (che però, non permette tutti i tipi di ricerche) e l'altro per esperti.

Un aspetto importante in questo àmbito è l'aiuto dato all'utente nell' uso della sintassi delle query e nei dettagli del tagging e del markup. Nei NUNC viene messo a disposizione dell'utente un sistema di autocomposizione (opera di Adriano Allora) che copre sia la sintassi del linguaggio regolare di CQP che la lista dei POS-tag utilizzati.

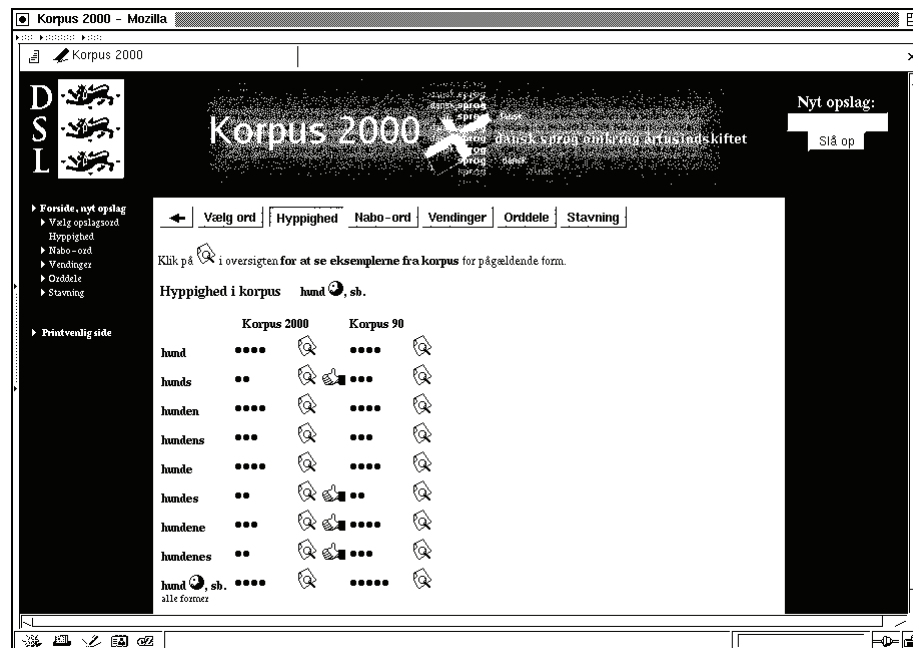
Nella versione semplificata dell'interfaccia del Bundestag Corpus e nell'interfaccia del Korpus 90 / Korpus 2000 danese gli elementi più essenziali della sintassi di CQP vengono inseriti automaticamente, e la schermata di ricerca si presenta così in modo abbastanza informale[25]: cfr. tavola 7.

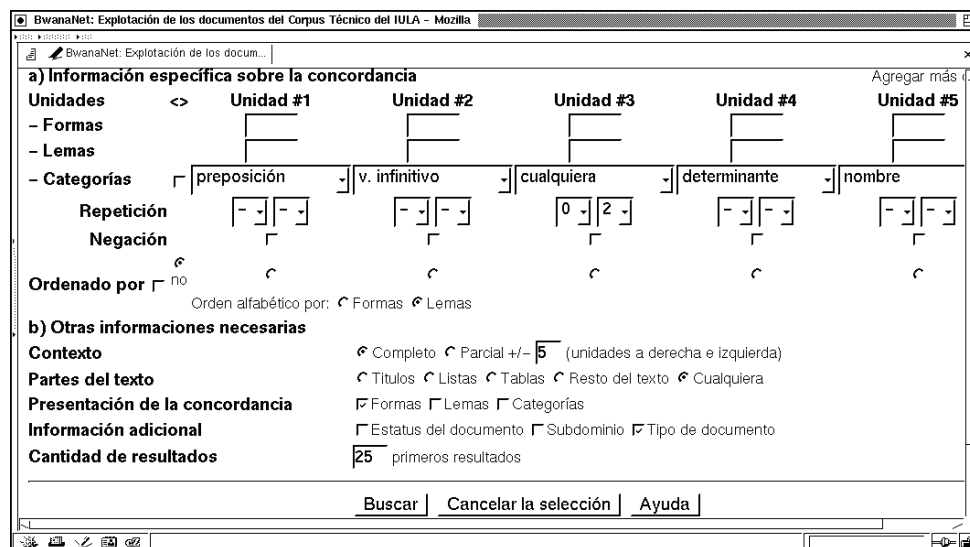
---

[23] La query, molto specifica, illustrata nella tavola è costruita per cogliere coppie *verbo* + *nome* in costruzioni causative al passivo con nome femminile plurale:

[15]
```
[lemma = "essere"] [word = "state"] [pos = "VER:pper"] [pos = "ADJ"]
{0,3} [pos = "NOM"];
```

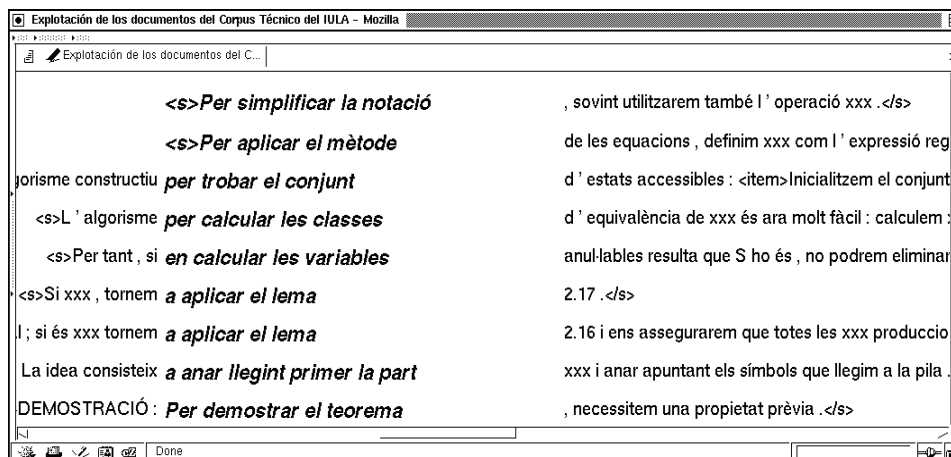
[24] In certo un senso, d'altra parte, questo già accade per i motori di ricerca come Altavista o Google.

[25] Accanto alla ricerca del lemma è anche possibile selezionare una delle forme flessive senza dover utilizzare la sintassi di CQP – semplificazione invero più fattibile per una lingua scandinava che per esempio per l'italiano.

Tav. 7. Interfaccia del Korpus 90 e Korpus 2000 danese: lista di forme flessive di *hund* (‘cane’) e frequenze nei due corpora.

In questo contesto è interessante la soluzione realizzata nell’interfaccia BwanaNet dello IU-LA di Barcellona.

Tav. 8. Query nell’interfaccia BwanaNet: preposizione seguita da un infinito, una serie di elementi arbitrari, un articolo ed un sostantivo. Corpus specialistico di informatica.

Accanto ad un’interfaccia semplice, infatti, è messa a disposizione dell’utente anche una versione avanzata in cui le espressioni regolari necessarie possono essere composte in modo

grafico: si possono cercare sequenze da 2 a 5 elementi (consecutivi o meno), ogni elemento può essere definito in termini di annotazioni, e per le etichette di categoria viene fornita una lista da cui selezionare: cfr. tav. 8.

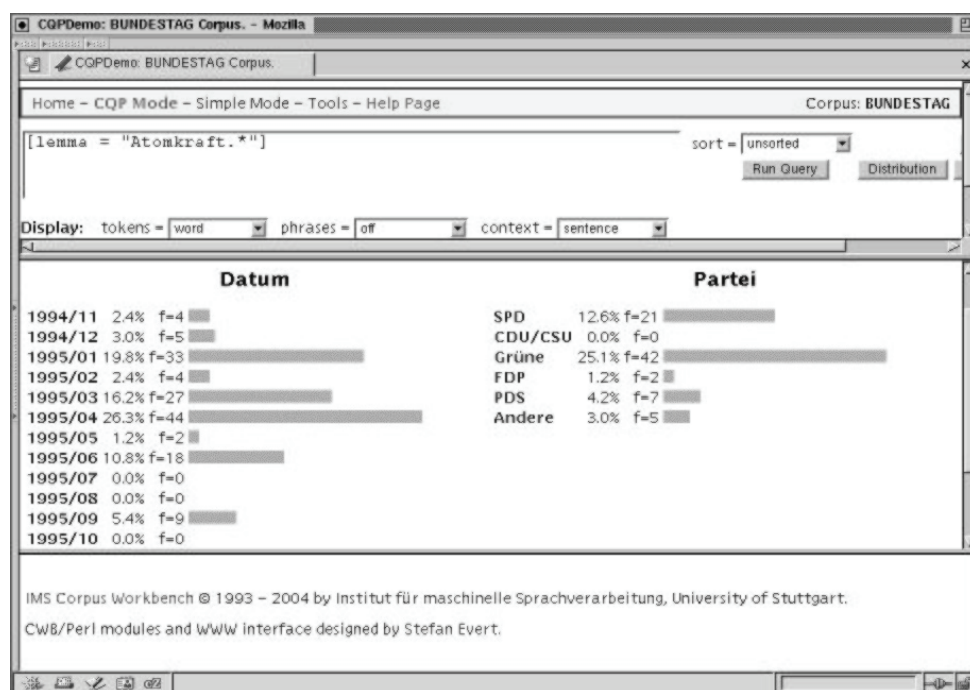
In BwanaNet i risultati possono poi essere organizzati in vario modo: cfr. tavola 9.

Tav. 9. Risultato della query della tavola 8 in forma di concordanza.

In generale, quindi, anche se il linguaggio CQP sembra in sé soddisfare la maggior parte delle necessità di una ricerca linguistica avanzata, è emersa chiaramente l'esigenza anche di un approccio semplificato (cfr. Hoffmann - Evert 2005). Due sono le possibili linee di realizzazione: da un lato un linguaggio meno formale (anche se meno potente) che permette di trasformare automaticamente una query verso il formato di CQP, e dall'altro un dialogo sequenziale, a diversi passi e diverse finestre (da percorrere una dopo l'altra)[26].

3.2 USO DI DIFFERENTI CORPORA SU UNA PIATTAFORMA COMUNE. Il CWB è stato inizialmente sviluppato per servire da piattaforma comune per i corpora dell'IMS. Uno degli obiettivi del progetto FIRB "L'italiano nella varietà dei testi" è similarmente quello di facilitare l'analisi parallela di fenomeni attraverso differenti tipi di testi e di corpora. L'interfaccia sviluppata a Torino è infatti utilizzata per i diversi corpora del FIRB; ed analogamente quella sviluppata a Barcellona offre un accesso comune a tutti i corpora disponibili allo IULA. Quel che però non è finora possibile con CQP, è l'interrogazione parallela di più corpora, ed il confronto incrociato dei risultati. Ovviamente, l'utente può far processare la stessa query su più corpora, ma solo sequenzialmente l'uno dopo l'altro.

Un aspetto che interessa molti linguisti è l'uso di strumenti diversi dal CQP con i risultati di query fatte con il CQP. A questo scopo, nel modo interattivo (in locale) esistono possibilità di salvare i vari dati su file (liste di risultati, liste di frequenza, e estratti di corpora). Ci sono due tipi di problemi legati alla messa a disposizione di tali strumenti sulla rete: tecnici e giuridici. Tecnicamente, un modello di comunicazione che consiste in una sola richiesta seguita da una sola risposta è ovviamente più facile da gestire (in termini sia di risorse di server che di banda).

[26] Quest'ultima via è quella seguita da BwanaNet, dove l'utente definisce la sua query incrementalmente, specificando, nell'ordine, la lingua, il subcorpus, il tipo di ricerca (semplice, standard od avanzata) ed, infine, la query stessa. Un tale dialogo, va però detto, se rende facile la definizione di parametri, è anche lungo da percorrere per un utente impaziente o che intenda compiere molte richieste di séguito.

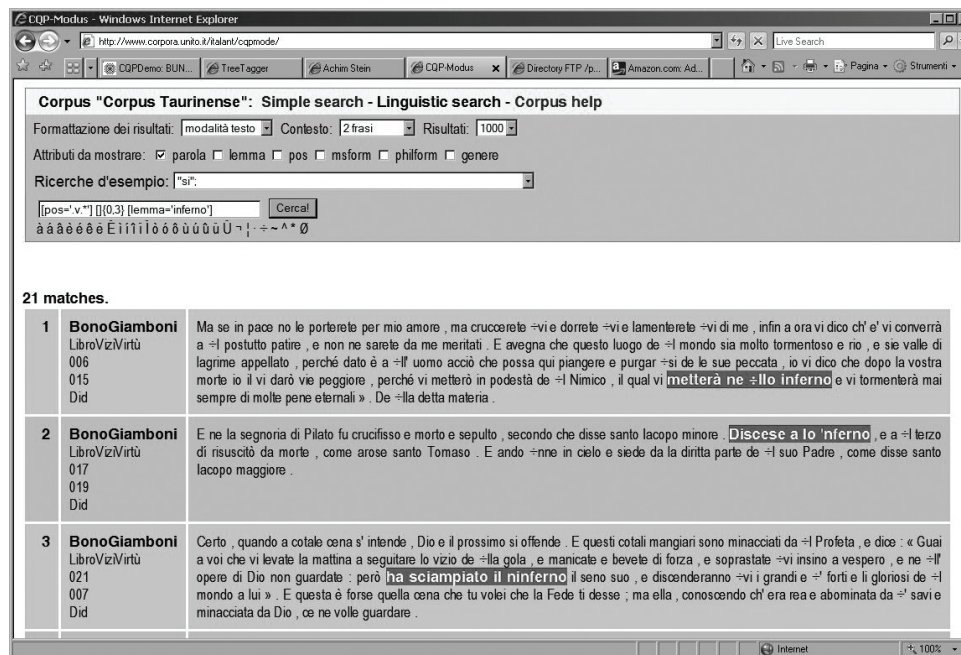
È comunque però possibile mettere a disposizione alcuni strumenti successivi, come ad esempio si è fatto all'IMS con le liste di frequenza nell'interfaccia del Bundestag Corpus: cfr. tavola 10.

Tav. 10. Interfaccia Bundestag (dimostrazione online di CQP all' IMS di Stoccarda) con indicazione della distribuzione per anni e partiti politici della parola *Atomkraft*[27] e dei suoi composti.

Dal punto di vista giuridico, poi, è talora discutibile se un istituto possa permettere a chiunque usa il sistema sulla rete di salvare parti anche importanti del corpus messo a disposizione. Ma di questo problema, centrale per il gruppo torinese di ricerca, si sono occupati altri dei contributi presenti in questo volume (cfr. Barbera ¶ 1 § 2.1, Allora - Barbera ¶ 5, Zanni ¶ 6 e Ciurcina - Ricolfi ¶ 7).

3.3 VISUALIZZAZIONE DEI RISULTATI. Sulla rete, come in locale in modo interattivo, la visualizzazione `kwic` ("key word in context"), cioè la concordanza tradizionale in ordine delle occorrenze nel corpus, è il modo più usato per presentare i risultati delle ricerche. Un esempio classico è il corpus portoghese AC/DC presente su Linguateca; una presentazione `kwic` si trova anche nella interfaccia BwanaNet (cfr. la tavola 9), con la parola chiave al centro e il contesto a destra ed a sinistra.

La visualizzazione dei tag (cfr. ad es. NUNC tavola 5), a richiesta dell'utente, è uno strumento utile per chi vuole analizzare in dettaglio il materiale estratto del sistema. Lo stesso sarebbe utile per i metadata, in quanto concerne tipi di testi. Anche per questo, l'interfaccia Bundestag, usata in forma leggermente modificata anche per il Corpus Taurinense, può servire da esempio.

---

[27] Dall'esame dei risultati di questa query emerge, tra l'altro, che la parola *Atomkraft* è preferita dagli ecologisti e dal partito socialdemocratico; in modo analogo si potrebbe constatare che *Kernkraft* ha una presenza "politica" specularmente contraria.

Tav. 11. Interfaccia Bundestag modificata per il Corpus Taurinense.

4. CONCLUSIONI. Il CWB è a nostra conoscenza l'unico sistema liberamente disponibile (cioè non proprietario) che possa manipolare corpora di grande estensione. In locale, è stato possibile lavorare con corpora di circa 300 milioni di parole, ed i NUNC generici italiani disponibili sulla rete contengono ciascuno più di 100 milioni di parole. Anche i corpora portoghesi vanno verso i 100 milioni di parole ciascuno. Per motivi tecnici non è al momento possibile lavorare con corpora di più di 300 milioni di parole, ragione per cui, ad esempio, il progetto nazionale ceco ha sviluppato uno strumento proprietario sulla base teorica del CQP. Per lavori con maggiori quantità di parole è per ora sempre possibile eseguire la stessa richiesta su differenti subcorpora e ricombinare i risultati in séguito.

Nella sua forma attuale, il CWB offre parecchie possibilità per la ricerca linguistica, interattiva ed automatizzata. Nel 2007, il CWB sarà messo a disposizione sotto contratto GPL (Gnu Public Licence), per permettere a tutti gli interessati di collaborarne allo sviluppo[28]. Le bozze delle specifiche per una nuova generazione di sistemi CQP per l'interrogazione e l'esplorazione di corpora si trovano in Hoffmann - Evert 2005; si può sperare che il CWB potrà essere utilizzato anche per la nuova frontiera della linguistica dei corpora, cioè l'uso del *web as a corpus* (cfr. Barbera - Corino - Onesti ¶ 3, § 1.5).

[28] L'autore della versione attuale di CQP, Stefan Evert, continuerà ad essere disponibile per ogni discussione.

BIBLIOGRAFIA.

ALLORA - BARBERA

¶ 5 Adriano Allora - Manuel Barbera, *Il problema legale dei corpora. Prime approssimazioni*, in questo volume, pp. 109-118.

BARBERA

2007 i.s. Manuel Barbera, *I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo*, in "Cuadernos de filología italiana" XIV (2007) 11-32, in corso di stampa.

¶ 1 Manuel Barbera, *Tra bmanuel.org e corpora.unito.it. Per la storia di un gruppo di ricerca*, in questo volume, pp. 3-20.

¶ 8 Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in questo volume, pp. 135-168.

BARBERA - CORINO - ONESTI

¶ 3 Manuel Barbera, Elisa Corino, Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

BRAUN - KOHN - MUKHERJEE

2006 *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, edited by Sabine Braun, Kurt Kohn and Joybrato Mukherjee, New York, Peter Lang, 2006 "English Corpus Linguistics" 3.

CHRIST et alii

1999 Oliver Christ - Bruno M[aximilian] Schulze - Anja Hofmann - Esther König, *The IMS Corpus Workbench: Corpus Query Processor (CQP). User's Manual*, Stuttgart, Institut für maschinelle Sprachverarbeitung, August 16, 1999 (CQP V2.2), documento disponibile online come file HTML (`http://www. ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/`), PS (`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/PS/cqpman.ps.gz`) o PDF (`http://www.ims.uni-stuttgart.de/projekte/CorpusWorknbench/CQPUserManual/PDF/cqpman.pdf`).

CHRIST - SCHULZE

1996 Oliver Christ - Bruno Maximilian Schulze, *CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora*, in FELDWEG - HINRICHS 1996; disponibile online alla pagina `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tue bingen.94.ps.gz`.

CIGNETTI

¶ 11 Luca Cignetti, *Alcune forme di polifonia testuale nei notiziari accademici di Athenaeum. Aspetti funzionali ed argomentativi*, in questo volume, pp. 199-207.

CIURCINA - RICOLFI

¶ 7 Marco Ciurcina - Marco Ricolfi, *Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora*, in questo volume, pp. 127-132.

FELDWEG - HINRICHS

1996 *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, herausgegeben von Helmut Feldweg und Erhard W. Hinrichs, Tübingen, Max Niemeyer Verlag, 1996 "Lexicographica. Series maior" 73.

GARSIDE - LEECH - MCENERY

1997 *Corpus Annotation. Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech and Anthony McEnery, New York, Longman, 1997.

HOFFMANN - EVERT
2006 Sebastian Hoffmann - Stefan Evert, *BNCweb (CQP edition): The Marriage of Two Corpus Tools*, in BRAUN - KOHN - MUKHERJEE 2006, pp. 177-195.

LEMNITZER - ZINSMEISTER
2006 Lothar Lemnitzer - Heike Zinsmeister, *Korpuslinguistik: eine Einführung*, Tübingen, Gunter Narr Verlag, 2006 "Narr Studienbücher".

MCENERY - XIAO - TONO
2006 Tony McEnery - Richard Xiao - Yukio Tono, *Corpus-Based Language Studies*, Abingdon, Routledge.

TOGNINI-BONELLI
2001 Elena Tognini-Bonelli, *Corpus Linguistics at Work*, Amsterdam - Philadelphia, John Benjamins Publishing Company, 2001 "Studies in Corpus Linguistics" 6.

ZANNI
¶ 6 Samantha Zanni, *Corpora elettronici e copyright. Lo stato legale della questione*, in questo volume, pp. 119-126.

CORPORA, STRUMENTI ED ISTITUZIONI DI RIFERIMENTO[29].

AltaVista	`http://www.altavista.com/`
Bundestag Corpus	`http://www.ims.uni-stuttgart.de/projekte/CQP Demos/Bundestag/frames-cqp.html`
corpora.unito.it	`http://www.corpora.unito.it/.`
Corpus Taurinense	`http://www.bmanuel.org/projects/ct-HOME.html`
CWB	`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`
CWB Users' Corner	`http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/UsersCorner.html`
EURAC	`http://www.eurac.edu/index_it`
GNU	`http://www.gnu.org`
Google	`http://www.google.it/`
IMS Stuttgart	`http://www.ims.uni-stuttgart.de/ims-home. html.en`
IULA Corpora	`http://www.iula.upf.es/corpus/corpusuk.htm`
IULA català	`http://bwananet.iula.upf.edu/bwananet1a.ca.htm`
KBTUO	`http://tekstlab.uio.no/Bosnian/Corpus.html`
Korpus 2000	`http://korpus.dsl.dk/korpus2000/indgang.php`
Korpus 90	`http://korpus.dsl.dk/e-resurser/k90_info.php?lang=dk`
LexAlp	`http://217.199.4.152:8080/general/lexalp/index.php`
Linguateca	`http://acdc.linguateca.pt/acesso/`

[29] Aggiornati al 15 marzo 2007.

MLCC W0023	`http://www.elda.org/catalogue/en/text/ W0023.html`
NUNC	`http://www.bmanuel.org/projects/ng-HOME.html`
OPUS	`http://logos.uio.no/opus/`
Parole	`http://spraakbanken.gu.se/parole/`
Tree Tagger	`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`
WebCorp	`http://www.webcorp.org.uk/`
WordNet	`http://wordnet.princeton.edu/`

ADRIANO ALLORA - MANUEL BARBERA[1]
(Università di Torino)

5. Il problema legale dei corpora. *Prime approssimazioni.*

"You boil it in sawdust: you salt it in glue:
You condense it with locusts and tape:
Still keeping one principal object in view –
To preserve its symmetrical shape."
Lewis Carrol, *The Hunting of the Snark*, 1876, Fit the Fifth, vv. 93-6.

0. PREMESSA**.** «The concept of a corpus which is in the public domain – available unconditionally for all users – does not so far exists» ammetteva, rassegnato, Geoffrey Leech nel 1991 (Leech 1991, p. 11), tracciando un quadro abbastanza sconfortante degli effetti della legge («that most slowly evolving of human institutions», *ibidem*) sulla ricerca; quadro purtroppo sostanzialmente rimasto inalterato fino ad oggi. I corpora che abbiamo distribuito su corpora.unito.it sono però, finalmente, quanto di più simile possibile a quelli di cui Leech quindici anni fa lamentava l'assenza. Come si è arrivati a questo importante risultato è raccontato e spiegato in questo contributo e nei due successivi.

Più precisamente, il presente testo, o meglio il suo nucleo iniziale scritto nell'estate 2002, voleva essere un primo avvicinamento al problema generale del copyright e del "copyleft"[2] ed al problema legale dei corpora in particolare: da un lato intendeva documentare lo status generale della questione nella comunità della linguistica dei corpora, e dall'altro chiarificare la posizione e gli intendimenti del nostro gruppo di ricerca, nato intorno al Cofin (oggi PRIN) sull'Italiano antico, organizzatosi nell'associazione *bmanuel.org*, nutrito dal FIRB e propagatosi grazie a *corpora.unito.it*. La speranza, naturalmente, era quella di riuscire a definire efficacemente la posizione legale dei nostri corpora (specie quelli attinenti al progetto FIRB), in relazione tanto al proprio ruolo di produttore-gestore di risorse, quanto al tipo di licenze da preparare per i propri utenti, e quanto al tipo di contratti (qualora ve ne dovessero essere) da stipulare con i propri fornitori di risorse testuali (case editrici, autori, istituzioni ecc.).

Questo è stato possibile grazie (1) ad una più stretta e funzionale definizione di cosa è un corpus (cfr. Barbera - Corino - Onesti ¶ 3), e, soprattutto, grazie (2) al provvidenziale incontro con l'avvocato Marco Ricolfi e la cellula italiana di Creative Commons (CC), che fortunatamente sono torinesi e legati alla Università di Torino. Il loro contributo (sostanziato nei due interventi successivi) è stato risolutivo; abbiamo comunque pensato che anche questi primi nostri appunti, originariamente un mero documento interno, possano servire a loro introduzione e contestualizzazione.

[1] I paragrafi 2.1 e 2.2 sono da attribuire ad Adriano Allora, gli altri a Manuel Barbera. In realtà, anche se la svolta di fare del problema legale un punto programmaticamente centrale di un progetto di ricerca linguistica è colpa da imputare a Manuel Barbera, le modalità e le proposte contenute in questo articolo sono in larga misura frutto di discussioni comuni, che videro variamente coinvolti tutti i membri del nostro gruppo di ricerca tra cui, oltre agli autori, in particolare Marco Tomatis.

[2] Una chiara guida su questo problema è ora disponibile in Wikipedia, alla pagina *Aiuto: diritto d'autore*, `http://it.wikipedia.org/wiki/Aiuto:Diritto_d%27autore`.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp.109-118.

1.1 La comunità della *Corpus linguistics* ed il problema legale. Il problema degli aspetti legali della preparazione e distribuzione dei corpora è uno dei più sentiti e "sofferti" dalla comunità tutta dei linguisti computazionali. E diciamo "sofferti" perché la legge, in questo settore, è stata finora avvertita soprattutto come un grave impedimento alla ricerca, che ostacola l'acquisizione dei materiali ad essa necessari, e limita la circolazione dei risultati da essa conseguiti, in nome di una oltranzistica ed antiliberale tutela del copyright (spesso ipostatizzata negli interessi di una bieca ed oscurantista "lobby dei copyright").

Che questo corrisponda o meno alla realtà, non è qui in discussione. Ma è comunque un fatto che quasi tutti gli adetpti di questa disciplina hanno dei corpora che custodiscono nel proprio cassetto, rigorosamente privati, cui ricorrono furtivamente e senza darne troppa pubblicità, in quanto costruiti con materiali di assai dubbia legalità. La situazione, in altri termini, è più quella pruriginosa[3] ed occulta dell'erotismo nell'epoca Vittoriana che non quella onesta e pubblica della moderna ricerca scientifica. L'esigenza di superare l'impasse dei copyright e di potere licenziare i propri corpora con modalità grosso modo analoghe a quelle GNU (cfr. infra), è infatti generalmente (e genericamente) reclamata da tutti; e sulla *Corpora List*, che è un poco il forum ufficiale della nostra comunità, il thread sui "legal aspects of compiling corpora" è stato particolarmente rovente per tutta l'estate 2002.

La frustrazione di scontrarsi con strumenti inadeguati contro diritti ormai troppo consolidati ha portato spesso più a tentativi di aggiramento delle leggi vigenti, che non a proposte positive. Si va da posizioni molto responsabili, come quelle prese da Geoffrey Sampson, che nella volontà di seguire un'impostazione da GNU pur nella consapevolezza che il corpus non è semplicemente un software, conclude ponendo al posto della licenza questo "disclaimer" (alquanto seccato verso gli aspetti legali) nella pagina da cui si possono scaricare i suoi corpora:

> *«So far as I am concerned, anyone is welcome to take copies of these resources and to use them for any purpose; and as far as I am able to check, I am legally entitled to make that offer. (If this is not legally watertight enough for you, you will have to go into the legalities yourself.) Naturally, if you do anything public with some of these materials, Sussex University and I would appreciate an acknowledgement (and, in the case of SUSANNE, CHRISTINE, and LUCY, so would the Economic and Social Research Council (UK), which sponsored their creation).»*
>
> (Sampson 2006).

In questo caso, le intenzioni sono lodevoli, ma resta il desiderio di strumenti legali adatti. L'opinione più frequentemente espressa (cfr. il thread menzionato sulla *Corpora List*), però, è molto spesso assai meno responsabile e si può ridurre a questa provocatoria affermazione: «se le fonti dei vostri testi non possono permettersi avvocati di grido non avrete contestazioni; evitate fonti dai bilanci troppo buoni!». Consiglio di indubbia validità pratica ma certo assai poco legale ...

1.2 La nostra posizione. In linea di massima, potremmo dire che ci troviamo (1) più che d'accordo sulle aspirazioni ideali, riassunte nel paragrafo precedente, ma certo (2) di meno sulla pratica.

Per quanto concerne il punto (1), anche il nostro intento generale è quello di conferire ai dati la maggior libertà di accesso e manipolazione possibili, secondo una strategia che ormai da circa un ventennio si sta affermando nel mondo dell'informatica, conformemente all'esigenza sempre più avvertita nella comunità internazionale della *Corpus linguistics*. Tale intento, va inoltre detto, risponde ad una logica di coerenza rispetto all'operato del nostro gruppo di ricerca a partire dall'esperienza del CT (*Corpus Taurinense*), legittimandone l'originalità e rilevanza

[3] Che poi, come è stato a volte salacemente detto, i più recenti "web corpora" se non appropriatamente filtrati rischino di essere prevalentemente pornografici, è un'altra questione ancora!

nel panorama della linguistica dei corpora: tra i principi che hanno informato lo sviluppo del CT, infatti, sono sempre stati presenti (cfr. soprattutto Barbera 2001) l'idea di riutilizzo delle risorse, il concetto di "linguistica ecologica", e la possibilità di condivisione e riapplicabilità di sistemi. Riproporre anche per il FIRB, e le nuove risorse in sviluppo, tale ottica di riutilizzo, estendendola anzi anche all'aspetto proprietario, risulta quindi una coerente evoluzione della nostra ricerca.

D'altro canto un simile approccio qualifica ulteriormente la *Corpus linguistics* e il NLP (*Natural Language Processing*) in generale, come discipline naturalmente votate alla mediazione tra il più dinamico ed innovativo campo dell'informatica e gli universi di solito più cauti e tradizionali dell'editoria e delle discipline umanistiche.

Per quanto, invece, concerne il punto (2), non credevamo (né crediamo) che la ricerca del sotterfugio e le strategie di evitamento in genere siano sempre un buon modo[4] per fare i conti con la legge. L'idea, al contrario, era di affrontare, una volta tanto, in positivo la questione legale e, anziché fuggirla, cercare di invocare la legge a nostro favore[5]: non reputandoci dei lestofanti o dei tagliagole, perché non avremmo potuto?

Questo gambitto, anche se legalmente ed eticamente sensato, era certo arrischiato, come non abbiamo tardato a rendercene conto, vuoi per la nostra ignoranza legale, vuoi perché la precisa natura legale di un corpus è risultata questione, oltre che non ovvia, comunque mai veramente affrontata: per quanto preparato dalla nostra accorta definizione di cosa sia un corpus (cfr. Barbera et alii ¶ 3 cit., § 4) e dai presenti scavi preliminari, il successo in ciò sarebbe stato impossibile senza il competente e perspicace apporto legale di Marco Ricolfi e della sua squadra di Creative Commons. Comunque, retrospettivamente, a FIRB quasi concluso, credo che l'avere voluto porre (e risolvere) la questione legale come una delle basi portanti di un progetto di linguistica dei corpora, anziché cercare di neutralizzarla od ignorarla, sia una delle novità più significative della nostra ricerca[6]. Non è infatti un caso, per limitarsi al solo italiano, che risorse liberamente disponibili siano sostanzialmente assenti: corpora di italiano di riferimento, come ad esempio il CORIS per l'italiano scritto, hanno l'indubbio vantaggio di essere particolarmente ben bilanciati e ben strutturati, ma sono altrettanto scarsamente utilizzabili per questioni di restrizioni[7] imposte da una gestione tradizionale del diritto d'autore.

2.1 BREVE INTRODUZIONE A GNU. La maggiore novità degli ultimi anni, dal punto di vista della cultura della ricerca e della condivisione delle risorse culturali, è stata l'ideazione del "free-software" e dell' "open-source", specificata in GNU, che "idealmente" anche noi (come tutti i linguisti di corpora) vorremmo, infatti, prendere a modello per la gestione dei nostri dati, contribuendone alla diffusione anche in àmbiti più vasti. Se per una sintesi più accurata della storia di GNU si può utilmente rimandare alla lettura del manuale *free* Medri 2001, data la non generale conoscenza di queste vicende in ambienti umanistici, non sarà inopportuno riassumerne i sommi capi.

GNU (il cui nome GNU fu scelto secondo una tradizione *hacker* come acronimo ricorsivo che significa *GNU's Not Unix*), è, propriamente, sia un progetto (economicamente produttivo, tra l'altro), sia un sistema (ormai perfettamente istituzionalizzato), sia una generale filosofia

---

[4] Per quanto assai consono all' (ab)usato costume italico, cui non vorremmo qui conformarci.

[5] Mossa che potrebbe anche ricordare il vecchio principio di strategia: se non puoi vincere il nemico, fàttene un alleato.

[6] L'unico appello esplicito (anche se inascoltato) in questo senso credo sia stato quello di De Santis 2001, pp. 127-130; precedentemente la presentazione più ragionevole del problema è probabilmente Atkins - Clear - Ostler 1992, p. 4.

[7] Contestabile soprattutto è la finestra massima di contesto, fissata arbitrariamente a 160 caratteri, senza badare a confini di parola o di frase. La molestia di ciò è evidente, come si vede ad es. in Onesti - Squartini ¶ 15, in questo volume.

(centrata intorno ai valori di libertà): il suo referente a livello planetario è la Free Software Foundation e l'atto giuridico di maggiore impatto nel quale GNU si identifica è la licenza GPL. GPL significa *General Public Licence* (GNU 1991) ed è un documento legale stilato dal fondatore di GNU, Richard Stallman, collaboratori e giuristi.

La logica di libertà, espressa da GNU, è perfettamente integrata ad una visione economica moderna: il software GNU, che è free software, *non è* necessariamente gratuito, come Stallman ha ripetutamente spiegato in diversi documenti. La nozione di *free*, nella sua concezione, non ha infatti nulla a che vedere con il prezzo del prodotto o con la sua ipotetica gratuità, interessa invece le libertà concesse agli utenti del prodotto. Egli indica i seguenti quattro punti nella definizione di un *free-software*: (a) l'utente ha la libertà di eseguire il programma per qualsiasi scopo; (b) l'utente ha la libertà di modificare il programma secondo i propri bisogni, e deve avere quindi accesso al codice sorgente ("open source") del programma; (c) l'utente ha la libertà di distribuire copie del programma, gratuitamente o dietro compenso; (d) l'utente ha la libertà di distribuire versioni modificate del programma, così che la comunità possa fruire dei miglioramenti apportati; (e) ciò che deriva da un progetto GNU deve essere anch'esso GNU. Questo insieme di caratteristiche viene di solito designato come *copyleft*, per contrapporlo al concetto di *copyright*.

La GPL riformula in termini giuridicamente opportuni i cinque punti sopra elencati. Esistono tuttavia anche altri tipi di licenza[8] che si riconoscono a vario titolo nel progetto GNU o nel concetto di free-software: alla GPL va, infatti, almeno affiancata la cosiddetta *Lesser GPL* (GNU 1999), la quale permette che software proprietario non GNU includa elementi GNU, come ad esempio la libreria C, o, più genericamente, che un free software possa esere incluso in un software non-free senza per questo diventare proprietario. La Berkeley Standard Distribution (BSD) e la Mozilla Public License (MPL), ad esempio, permettono non solo che il free-software da loro tutelato venga introdotto in software proprietario, ma anche che le modifiche apportate al free-software possano essere mantenute private; e, ancora, la Netscape Public License (NPL) contiene alcuni privilegi esclusivi dell'azienda Netscape.

È utile sapere, poi, che la GNU ha approntato, accanto alla GPL per i software, anche una licenza per testi, la cosiddetta GNU FDL "GNU Free Documentation License" (GNU 2002). I testi primo oggetto del suo interesse (i manuali dei programmi) sono particolari, è vero, ma resta il fatto che presentano la maggior parte dei problemi legali di qualsiasi altro testo: gli autori, infatti, dicono esplicitamente nel preambolo: «We have designed this License in order to use it for manuals for free software [...]. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a print». Infatti sulla FDL è nata Wikipedia, la "free encyclopedia", la cui recente crescita (qualitativa e quantitativa) è una conferma della potenzialità culturale dell'orientamento GNU.

La dicotomia radicale, comunque, tra software (allora GNU) e non software (ed allora non GNU) non si pone quindi più in questi termini, e di ciò potrebbero giovarsi anche oggetti più ambigui situati tra questi due poli, come appunto i corpora.

L'impostazione del progetto GNU si è rivelata vincente dal punto di vista dei risultati scientifici prodotti ed anche dal punto di vista economico: il fatto che non sia stata adottata anche in altri settori è, quindi, probabilmente più una questione di reattività o diffidenza dei settori in questione, che non una effettiva previsione economica negativa. Il proposito così spesso espresso dai linguisti dei corpora di spostarsi sotto GNU conferma, semmai, come gli studiosi di *corpus linguistics* e NLP siano nella posizione privilegiata di ambire a fare una mediazione fra il mondo dell'informatica, quello della linguistica, e quello delle realtà economiche più dinamiche.

---

[8] Fondamentale, nel nostro caso, come vedremo in Zanni ¶ 6 e Ciurcina - Ricolfi ¶ 7 oltre in questo volume, è, al di là della FDL, la licenza *Attribution* di Creative Commons (CC 2002).

2.2. I GRANDI DISTRIBUTORI DI CORPORA: STRATEGIE E PROBLEMI. Ma, di fatto, quale è la situazione più normale e diffusa circa la reperibilità dei corpora attualmente diponibili, specie sulla rete? Si può dire che la regola generale applicata per i corpora disponibili in rete preveda il semplice copyright, con l'eccezione dei lavori di ricerca non commerciali: in questo modo i responsabili delle varie istituzioni pensano, un po' semplicisticamente, di proteggere i diritti delle loro fonti testuali, prevalentemente case editrici e giornali.

Per esempio, l'ACL (Association for Computational Linguistics), in merito alla DCI (Data Collection Initiative[9]), racconta che quel progetto fu avviato "to oversee the acquisition and preparation of a large text corpus to be made available for scientific research at cost and without royalties"; nell'ACL/DCI User Agreement viene chiesto di sottoscrivere l'impegno ad usare il corpus a soli fini di ricerca, ed in questo modo la ACL si sottrae al problema del copyright.

Il grande distributore ELRA (European Language Resource Association), che richiede una quota di associazione oltre al prezzo per l'accesso ai singoli corpora[10], prevede tre tipi di licenza (accademica, commerciale e di prova, valida per tre mesi) e anche se nel sito dell'associazione non si scende mai nel dettaglio riguardo al tipo di uso che si può concretamente fare dei corpora acquistati, è naturale l'accostamento dei primi due tipi alla consueta opposizione: usi non commerciali ed usi commerciali (con preventivo contratto con le fonti dirette).

L'ELAN (European Language Activity Network) richiede l'adesione ad una sorta di "comunità virtuale" i cui utenti possono accedere ai corpora disponibili ma, ancora una volta, avvertiti della possibilità di incorrere in reati contro il copyright in un certo numero di casi, ossia: quando, anche per usi non commerciali, vengano estratte dai corpora citazioni di più di 500 caratteri, oppure quando, per usi commerciali, le questioni relative ai diritti non vengano discusse anticipatamente con i singoli possessori dei medesimi.

La licenza di TRACTOR, l'importante distributore di corpora connesso alla TELRI (Trans-European Language Resources Infrastructure), non si discosta da questo standard, con la sola eccezione del limite dei 500 caratteri che non esiste (facendo sorgere seri dubbi sulla sua accettabilità, da parte degli utenti, o sulla sua efficacia, da parte dei fornitori).

L'OTA (Oxford Text Archive), poi, concede diritti di utilizzo a scopo non-commerciale, ma esclude completamente l'uso per scopi commerciali come anche la riproduzione senza il consenso di chi ha inserito nell'archivio il testo usato e/o riprodotto.

Se tutte quelle finora riferite sono iniziative comunque tra le più serie ed importanti a livello internazionale, si trovano poi anche imprese più furbesche, come un'organizzazione dall'ingannevole nome "Open Language Archives Community", la quale tuttavia specifica bene che «Open does not mean that users are free to do whatever they like with the metadata, nor does it mean that the described language resources are openly available». *Open*, insomma, starebbe solo a significare che gli iscritti possono liberamente guardare ed aggiungere qualsiasi "archivio".

Un esame anche sommario dei tipi di licenze di solito praticate, evidenzia soprattutto due punti legali quasi sempre sollevati, più al fine di tutelare i fornitori di dati che non di promuovere la fruizione delle risorse: (1) il timore di sviluppi commerciali e (2) la limitazione all'uso dei corpora. Questi due punti, in effetti, sembrano asimmetricamente tutelare i (presunti) interessi delle case editrici[11] e ledere le basilari possibilità d'azione degli utenti. Pur proteggendo i diritti sui testi originali, andrebbe invece fatto salvo il principio che i corpora debbano essere condivisibili liberamente, in prima istanza anche gratuitamente, e poter poi diventare parte di altri prodotti, od elemento dello sviluppo di altri prodotti, anche commerciali, dimodocché il ritorno economico costituisca un volano alla ricerca medesima.. Dalle misure cautelative finora

---

[9] La DCI è diventata recentemente partner del motore di ricerca Google con lo scopo di rendere pubblico il corpus di oltre mille miliardi di parole raccolto dal noto motore di ricerca (cfr. Google N-grams Corpus)
[10] E che nei limiti del presente testo prenderemo ad esempio per tutti i grandi distributori di corpora a pagamento.
[11] Ricordate la famigerata "lobby dei copyright" che si menzionava all'inizio?

adottate, invece, non può in effetti venire, a nostro parere, né una promozione della ricerca né un reale utile degli enti fornitori di testi, specie se questi sono quegli enti commerciali, come le case editrici, che si vorrebbe invece favorire.

3. VERSO UNA SOLUZIONE. Iniziamo ora ad esaminare i pro ed i contro delle diverse soluzioni che si sono date al problema, giungendo a proporne anche una possibile nostra.

3.1 LE VIE USATE. Le soluzioni più interessanti sul tappeto sono essenzialmente due.

La tattica più semplice **(1)** è certo quella (fatta propria ad es. dall'ACL/DCI) di conformarsi alla tradizionale incapacità giuridica propria delle istituzioni universitarie di fornire e/o vendere dati per applicazioni commerciali; in altre parole la "soluzione" consisterebbe nel limitare rigidamente l'uso dei corpora alle sole finalità non commerciali, senza scopo di lucro. Tale soluzione comporta, però, almeno due ordini di controindicazioni: (a) paradossalmente seguendo questo schema in modo rigoroso una casa editrice, poniamo, potrebbe non essere più in condizione di riutilizzare (per i propri scopi commerciali) i corpora costruiti a partire dai dati da essa forniti, perché il loro utilizzo è ormai vincolato all'assenza di fini di lucro[12]; (b) la possibilità di utilizzare i corpora anche per sviluppare applicazioni commerciali potrebbe essere, come accennato, un volano per il successo di utilizzo del corpus stesso e, quindi, l'autofinanziamento di nuova ricerca, che così non graverebbe (o graverebbe di meno) sulle sempre più pericolanti casse della ricerca pubblica.

Una seconda soluzione **(2)** sarebbe quella di trovare un discrimine legale tra *riproduzione* ed *utilizzo* (vietata la prima, ammesso il secondo) dei testi sorgente, ossia del corpus non etichettato. La negata *riproduzione* potrebbe tranquillizzare, ad esempio, una casa editrice in merito alla possibilità che i testi che fornisce vengano riprodotti *ipso facto* in una nuova edizione da parte di un qualche ipotetico editore pirata, mentre *l'utilizzo* garantirebbe la possibilità di impiego anche per scopi commerciali del corpus (purchè non per un'edizione clone di quella di partenza, coperta da diritto e proprietaria, fornita dalla casa editrice al costruttore del corpus), facendo salvi i diritti della casa editrice. Di fatto finora ciò è equivalso a restringere la disponibilità dei corpora a contesti quantitativamente determinati: il CORIS, ad esempio, come s'è detto, limita la riproduzione a contesti di 160 caratteri e l'ELAN a 500, e questa strategia in generale pare essere la norma. Tale pratica è fortemente dannosa per la ricerca linguistica[13], e prova ne è, infatti, lo scarso utilizzo da parte dei linguisti di risorse che fanno ricorso a questi tipi di limitazione. Inoltre, definire giuridicamente il discrimine tra *riproduzione* ed *utilizzo* senza fare ricorso a limiti estrinseci (castranti per la ricerca, con produzione di risorse di scarso utilizzo, quindi anche con poco ritorno di immagine per l'ente fornitore, e pertanto da evitare nell'interesse di tutti) è piuttosto arduo. Forse non è dunque questa la strada da percorrere.

3.2 UNA NUOVA PROPOSTA. Una terza possibile maniera **(3)** di risolvere il problema, non completamente alternativa alla seconda, quanto piuttosto ad essa complementare, potrebbe essere quella di precisare in chiari termini, sulla scorta della nostra specifica definizione di corpus (cfr. Barbera - Corino - Onesti ¶ 3, § 4), la differenza tra i semplici testi (che chiameremo, per semplificare, *T*, il corpus nudo) così come forniti, ad esempio, da una casa editrice ed il testo

---

[12] Esistono precedenti di ciò; ma esistono, d'altro canto, anche consolidati ed efficaci espedienti per evitarlo. L'obiezione che bisognerebbe scartare questa ipotesi per in ogni modo garantire che la casa editrice (od altro ente commerciale fornitore di testi) un giorno possa utilizzare il corpus per qualsiasi applicazione anche a scopo commerciale, è pertanto più *de jure* che *de facto*.

[13] Ricerche di tipo testuale, semantico o pragmatico, come alcune di quelle qui presentate nel prosieguo (cfr. ad es. Carmello ¶ 21, Cignetti ¶ 11, Ferrari - Mandelli ¶ 10, ecc.), sarebbero anzi del tutto impossibili.

arricchito di markup, tokenizzazione ed altri possibili tag, che costituisce il corpus (che chiameremo *T+n*) concretamente prodotto.

In questo modo si potrebbe efficacemente differenziare anche legalmente tra *T+n* (il testo sottoposto a lavorazione più tutte le etichette), in uso pieno ed assoluto secondo i criteri dello standard GNU, e *T*, il testo nudo fornito dalla casa editrice, che continuerebbe ad essere sottoposto alla normale tutela legale. In altre parole il corpus (*T+n*) sarebbe in regime di *copyleft*, mentre i testi di partenza resterebbero in *copyright* alla casa editrice, grazie alla libertà dei prodotti GNU di comprendere anche parti proprietarie. L'utilizzatore, in tale prospettiva, potrebbe quindi fare l'uso che preferisce, anche commerciale, di qualsiasi parte di *T+n*, ma qualora lo riconverta a *T*, rifiuterebbe di fatto la licenza GNU di *T+n* e ritornebbe sottoposto al diritto ordinario che tutela la proprietà di *T*, in base al quale potrebbe essere normalmente perseguito.

In questa, terza, opzione, i problemi che potrebbero sorgere dal punto di vista del costruttore / distributore del corpus per la "minore tutela" della sua opera sarebbero perlopiù apparenti[14], e così lo sarebbero anche quelli dei detentori di diritti dei testi iniziali. Mettiamo pure in un caso di riutilizzo virtuoso: quello che avverrebbe non è in realtà una *riproduzione* del semplice testo originario (*T*), ma un *utilizzo* del testo originario con parte del markup aggiunto dal primo corpus (*T+1*), per produrre un terzo corpus (*T+1+n*), che è perlomeno la somma di *T* più quanto verrà aggiunto in un secondo momento. Infatti, anche qualora il riutilizzatore in questione rinuncasse ad una o più fascie di annotazione del corpus precedente, come ad esempio quella morfosintattica, non per questo rinuncerebbe al template generale del markup impiegato, né alla tokenizzazione che era stata applicata al testo; testo che sarebbe comunque ancora sempre un *T+n* differente dal *T* testo nudo fornito dall'editrice. Tutto ciò continuerebbe pertanto a lasciare, da un lato, intatto il diritto alla riproduzione a scopo commerciale che la casa editrice esercita sulla propria opera, così come il diritto a rivalersi di qualsiasi riproduzione pirata di *T*; e, dall'altro lato, a non limitare la ricerca possibile a partire dal corpus *T+n*, consentendo (anche economicamente) la generazione di una serie virtualmente illimitata di *(T+n)+n+n*...

Anche in quest'ultima più promettente prospettiva, però, una elaborazione legale competente ed efficace ci appariva comunque assolutamente necessaria. Per tacere di minori problemi, potrebbe, ad esempio, parere in conflitto con la normativa vigente (legge sul diritto d'autore modificata dal D.lgs. 6 maggio 1999, n. 169 attuativo della direttiva 96/9/CE relativa alla tutela giuridica delle banche di dati), in cui i corpora ricadono a pieno titolo in quanto raccolte di «opere, dati o altri elementi indipendenti sistematicamente o metodicamente disposti ed individualmente accessibili mediante mezzi elettronici o in altro modo».

I risolutivi contributi Zanni ¶ 6 e Ciurcina - Ricolfi ¶ 7 (qui oltre) dimostreranno che ciò non è ed apriranno anzi, su queste basi, le porte ad una efficace regolamentazione contrattuale anche per il nostro tormentato settore.

[14] Anche al di là del fatto che, in una prospettiva GNU la possibilità stessa che un'altra istituzione decida di etichettare un nostro stesso corpus con un diverso fine, auspicabilmente nella conformità alle metodologie informatiche da noi applicate, non può che rappresentare un arricchimento del corpus, *per sé* desiderabile, per noi proficuo (in quanto potrebbero venirci richiesti servizi o richieste di sviluppo), e vantaggioso anche per l'originario fornitore dei materiali nudi, che vede in tal modo pubblicizzati tanto i propri prodotti quanto la propria sensibilità alla ricerca.

BIBLIOGRAFIA.

AIJMER - ALTENBERG
1991 *English Corpus Linguistics. Studies in Honour of Jan Svartvik,* edited by Karin Aijmer e Bengt Altenberg, London - New York, Longman, 1991.

ATKINS - CLEAR - OSTLER
1992 Sue Atkins - Jeremy Clear - Nicholas Ostler, *Corpus Design Criteria*, in "Literary and Linguistic Computing" VII (1992)[1] 1-16.

BARBERA
2001 Manuel Barbera, *From EAGLES to CT Tagging: a Case for Re-usability of Resources*, in RAYSON et alii 2001, pp. 40-44.

BARBERA - CORINO - ONESTI
¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

CARMELLO
¶ 21 Marco Carmello, *"Dovere" deontico e "dovere" anankastico fra semantica e pragmatica. Una ricerca corpus-based*, in questo volume, pp. 347-362.

CC
2002 Creative Commons, *Attribution License*, versione 2.0, [16 dicembre 2002], disponibile online alla pagina `http://creativecommons.org/licenses/by/2.0/legalcode`.

CIGNETTI
¶ 11 Luca Cignetti, *Alcune forme di polifonia testuale nei notiziari accademici di Athenaeum. Aspetti funzionali ed argomentativi*, in questo volume, pp. 199-207.

CIURCINA - RICOLFI
¶ 7 Marco Ciurcina - Marco Ricolfi, *Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora*, in questo volume, pp. 127-132.

DE SANTIS
2000 Cristina De Santis, *Isole e tesori. Navigare alla ricerca di risorse per la costruzione di un* corpus *di italiano scritto*, in ROSSINI FAVRETTI 2000, pp. 121-131.

FERRARI - MANDELLI
¶ 10 Angela Ferrari - Magda Mandelli, *Note sull'impiego dei connettivi nei notiziari accademici del corpus Athenaeum. Aspetti quantitativi e qualitativi*, in questo volume, pp. 183-198.

GNU
1991 *GNU General Public License (GPL)*, versione 2, giugno 1991, disponibile online alla pagina `http://www.gnu.org/licenses/gpl.html`.
1999 *GNU Lesser General Public License (Lesser GPL)*, versione 2.1, febbraio 1999, disponibile online alla pagina `http://www.gnu.org/licenses/lgpl.html`.
2002 *GNU Free Documentation License (FDL)*, versione 1.2, novembre 2002, disponibile online alla pagina `http://www.gnu.org/copyleft/fdl. html`.

LEECH
1991 Geoffrey Leech, *The state of the art in corpus linguistics,* in AIJMER - ALTENBERG 1991, pp. 8-29.

MEDRI
2001 Daniele Medri, *Linux Facile*, versione 5.0, 10 giugno 2001, online alla pagina http://www.linuxfacile.org; prima edizione cartacea: Bologna, Silicon Graphics Italia, 2000; poi anche Milano, Lumina SAS, e Milano, Systems SRL (allegato alla rivista "Inter.net").

ONESTI - SQUARTINI
¶ 15 Cristina Onesti - Mario Squartini, *"Tutta una serie di". Lo studio di un pattern sintagmatico e del suo statuto grammaticale*, in questo volume, pp. 271-284.

RAYSON et alii
2001 *Proceedings of the Corpus Linguistics 2001 Conference. Lancaster University 29 March - 2 April 2001*, edited by Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja, Lancaster, University Center for Computer Corpus Research on Language, 2001 "UCREL Technical Paper" 13.

ROSSINI FAVRETTI
2000 *Linguistica e informatica.* Corpora*, Multimedialità e percorsi di apprendimento*, a cura di Rema Rossini Favretti, Roma, Bulzoni Editore, 2000.

SAMPSON
2006 Geoffrey Sampson, *Downloadable Research Resources*, web page, ultima versione 30 settembre 2006, http://www.grsampson.net/Resources.html.

ZANNI
¶ 6 Samantha Zanni, *Corpora elettronici e copyright. Lo stato legale della questione*, in questo volume, pp. 119-126.

CORPORA, GESTORI DI CORPORA ED ALTRI SITI DI RIFERIMENTO[15].

ACL/DCI	http://www.ldc.upenn.edu/Catalog/
bmanuel.org	http://www.bmanuel.org/projects/index.html
BSD	http://www.opensource.org/licenses/bsd-license.html
CC	http://creativecommons.org/
CC-it	http://it.creativecommons.org/
CORIS	http://corpora.dslo.unibo.it/coris_ita.html
Corpora List	http://listserv.linguistlist.org/archives/corpora.html
corpora.unito.it	http://www.corpora.unito.it/
Corpus Taurinense	http://www.bmanuel.org/projects/ct-HOME.html
ELAN	http://nl2.ijs.si/index-bi.html
ELRA	www.elra.info
Free Software Foundation → GNU	
G. Sampson h.p.	http://www.grsampson.net/

[15] I link sono stati aggiornati al 18/02/2007.

GNU	www.gnu.org www.fsf.org
Google N-grams Corpus	http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
MPL	http://www.mozilla.org/MPL/
Netscape	http://www.netscape.com/
NPL	http://www.mozilla.org/MPL/NPL-1.0.html
OTA	http://ota.ahds.ac.uk/
TEI	http://www.tei-c.org/
TELRI	http://nl.ijs.si/telri/
Wikipedia	http://en.wikipedia.org/wiki/Main_Page http://it.wikipedia.org/wiki/Pagina_principale

SAMANTHA ZANNI
(Università di Torino)

6. Corpora elettronici e copyright. *Lo status legale della questione.*

0.1 PREMESSA GENERALE**.** Nell'intento di soddisfare le esigenze, affermatesi nella comunità internazionale delle discipline umanistiche, di maggiore libertà di accesso e manipolazione possibili di dati testuali, si sta diffondendo l'utilizzo e lo sviluppo di varie risorse linguistiche, quali i "corpora", di solito intesi, nella loro accezione comune, come una raccolta di testi autentici e ricorrenti nell'uso, in formato elettronico, selezionati come rappresentativi (per es.) dell'italiano corrente. Tale definizione, che già coprirebbe molti "usi" del termine "corpus" va comunque ulteriormente specificata, nel senso di Barbera - Corino - Onesti ¶ 3 (cfr. soprattutto § 4) qui sopra, per ricoprire l'uso particolare dei "linguisti computazionali"[1] che si occupano di corpora.

Per "corpus" bisogna propriamente intendere, quindi, una raccolta di testi autentici ed in formato elettronico trattati in modo da essere gestibili ed interrogabili informaticamente e sui quali vengono applicati modelli e tecniche computazionali (tokenizzazione, markup, tagging ecc.).

Le operazioni di "tokenizzazione", "markup" e "tagging" deformano il testo originario, costituiscono la componente informatica del testo ed hanno una propria dignità creativa, attribuendo al testo una sorta di natura "intrinsecamente" informatica cui si affianca l'accessibilità (anch'essa) informatica conferita dal supporto tecnico (Web o CD-Rom) attraverso il quale il corpus viene messo a disposizione degli utenti[2].

Il linguista computazionale analizza e scompone il testo in modo da poter costruire una sequenza di operazioni semplici (istruzioni) che offrono una soluzione di lettura del testo medesimo. Il passo tecnico successivo (che può anche essere solo eventuale), è la traduzione di questa sequenza di istruzioni in un vero e proprio programma, o meglio linguaggio di programmazione, come nell'esempio (cfr. *supra* Barbera - Corino - Onesti ¶ 3, cit., Tav. 5) della traduzione nel formalismo (linguaggio) CQP[3] del testo tokenizzato e markuppato del Corpus Taurinense. Un altro esempio di modello computazionale che si è inserito nella costruzione di un programma potrebbe essere l'analizzatore sintattico, o "parser", che è un programma che legge una frase (o un testo, una frase alla volta) e, consultando una grammatica opportunamente definita e scritta in un formalismo specifico, restituisce in output la struttura sintattica della frase.

Al di là di tali ipotesi di correlazione, per così dire, "automatica" tra linguaggio e calcolatore, la manipolazione "manuale" o "semi-automatica" od anche "automatica"del testo resa possibile dalla linguistica computazionale non crea un linguaggio di programmazione in senso proprio e non riconduce il "corpus" alla disciplina apprestata dalla normativa vigente con riguardo ai programmi per elaboratore.

0.2 PREMESSA PARTICOLARE**.** L'Università di Torino, nelle persone di Carla Marello e Manuel Barbera, ha evidenziato l'interesse alla pubblicazione on line di materiali "senza restrizioni di diritti". In particolare, essa intende realizzare un progetto finalizzato alla libertà di

[1] La linguistica computazionale è l'area disciplinare che si fonda su di una relazione tra studio teorico della lingua e calcolatori.
[2] Per i vari tipi di tagging e di markup cfr. *supra* Barbera - Corino - Onesti ¶ 3. Il markup, aggiungiamo, può anche considerarsi come un metodo unificante di rappresentazione delle varie etichette (tag) di un testo.
[3] Il linguaggio CQP è più dettagliatamente illustrato nel contributo di Heid ¶ 4, sopra in questo volume.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp.119-126.

accesso, di utilizzo e di sviluppo di taluni "corpora" (nel prosieguo denominati "Corpora") da parte degli utenti, messi a disposizione via Web o su CD-Rom. Viene chiesto quindi di individuare gli strumenti giuridici maggiormente idonei al raggiungimento dello scopo sopra delineato.

Dalla lettura delle note programmatiche per la definizione dell'assetto legale dei Corpora, fornite dal gruppo di lavoro (qui sostanzialmente riprodotte come Allora - Barbera ¶ 5), emergerebbe quanto segue: (a) il progetto dei Corpora è ideologicamente conforme all'iniziativa GNU[4], ideatrice del *"free software"* e dell' *"open source"*; (b) il gruppo di lavoro ha espresso il desiderio di poter utilizzare, per la concessione in uso dei dati contenuti nei Corpora, licenze tipo "GPL" ovvero "Lesser GPL" (messe a disposizione da GNU su web) da siglarsi con gli utenti, senza pregiudizio dei diritti spettanti sulle opere originarie in capo ai fornitori delle risorse testuali – case editrici, autori, istituzioni ecc.– (cfr. Allora - Barbera cit. ¶ 5).

0.3 IL PRESENTE CONTRIBUTO. Alla luce di queste precisazioni, si ritiene che possano essere formulate talune considerazioni che qui di seguito vengono esposte articolatamente.

Si ritiene inoltre che, in generale, le bozze di contratto di licenza "Creative Commons" possano soddisfare le esigenze espresse dalla comunità torinese della *Corpus linguistics* (libertà di distribuzione del corpus, libertà di modifica del corpus, libertà di distribuzione del corpus modificato, ecc.), anche se taluni argomenti potranno essere ancora approfonditi e definiti

1.1 CORPUS "OPERA DERIVATA" ED "OPERA COLLETTIVA". Il "corpus" si costituisce mediante l'apposizione, su di una base dati testuale, di una serie di notazioni proprie della linguistica computazionale. Esso rappresenta quindi una "rielaborazione" di una o più opere aventi carattere di creazioni autonome, costituendo, ciascuna "rielaborazione" un'"opera derivata" ai sensi della disciplina apprestata dalla legge sul diritto d'autore.

In quanto raccolta di opere ("derivate") aventi carattere di creazioni autonome, il "corpus" risulta altresì qualificabile come "opera collettiva", sottostando, conseguentemente, alla disciplina apprestata dalle disposizioni di cui agli artt. 3, 7 e 38 l.a.

1.2 CORPUS "BANCA DI DATI". Il "corpus" è qualificabile inoltre come "banca di dati", in quanto identificabile come raccolta di opere (derivate) «indipendenti, sistematicamente o metodicamente disposte ed individualmente accessibili, dotate di creatività nella scelta ovvero nella disposizione dei materiali», e quindi disciplinata dagli artt. 64 *quinquies* e *sexies*, a protezione della "scelta e della disposizione del materiale raccolto".

1.3 CORPUS TUTELATO DAL DIRITTO "SUI GENERIS". Qualora *il conseguimento, la verifica e la presentazione del contenuto informativo della banca di dati* richiedessero un *investimento rilevante* (anche solo in termini di risorse intellettuali), al costitutore della banca di dati potrebbe spettare altresì il diritto quindicinale "*sui generis"* (di cui all'art. 102 *bis)* e dunque una protezione afferente il *contenuto informativo* del corpus (con conseguente diritto di vietare / consentire operazioni di reimpiego od estrazione della parte o della totalità del contenuto del "corpus"[5]).

[4] Cfr. il sito Gnu e Linux Facile, e qui il § 2.1 di Allora - Barbera ¶ 5.

[5] Si ricorda che per *estrazione* il disposto normativo intende «il trasferimento permanente o temporaneo della totalità o di una parte sostanziale del contenuto di una banca di dati su un altro supporto con qualsiasi mezzo o in qualsivoglia forma», e per *reimpiego* «qualsivoglia forma di messa a disposizione del pubblico della totalità o di una parte sostanziale del contenuto della banca di dati mediante distribuzione di copie, noleggio, trasmissione effettuata con qualsivoglia mezzo e in qualsiasi forma».

2.1 CREAZIONE E RIPRODUZIONE DEL CORPUS - NECESSITÀ DEL CONSENSO DELL'AUTORE DEL SINGOLO CONTRIBUTO O SUO AVENTE CAUSA. Dovranno prestare il proprio consenso all'elaborazione delle proprie opere attraverso i modelli e le tecniche proprie della linguistica computazionale i titolari dei diritti sui singoli contributi (autori o aventi causa).

Ciò non è però necessario nel caso in cui l'opera non sia più tutelata dal diritto d'autore per essere cessati i termini di questa (70 anni dalla morte dell'autore) previsti dall'art. 25 L. 633/41.

Si segnala inoltre che, ai sensi dell'art. 5 della stessa L. 633/41, le disposizioni della Legge «non si applicano ai testi degli atti ufficiali dello Stato e delle Amministrazioni pubbliche, sia italiane che straniere».

2.2 ATTRIBUZIONE DEI DIRITTI PATRIMONIALI DI SFRUTTAMENTO DEL CORPUS – NECESSITÀ DI CONSENSO DEGLI ELABORATORI E DELL'ORGANIZZATORE DEL CORPUS. Anche un ente pubblico può essere titolare dei diritti d'autore su di un "corpus" a condizione che (j) non sussistano norme interne che impediscano l'acquisizione di diritti patrimoniali su di un'opera dell'ingegno e (ij) ferma restando la necessità che siano stipulati con gli elaboratori del "corpus", ovvero di colui che dirige ed organizza il corpus, pattuizioni contrattuali (di lavoro subordinato o autonomo) che sanciscano l'attribuzione in capo all'ente dei relativi diritti di utilizzazione. In questo caso non è esclusa la fattibilità di uno "spin-off" creato *ad hoc* per l'attribuzione dei diritti relativi ai "Corpora".

Nel caso dei corpora prodotti dal gruppo torinese, si è preferito (per problemi di organizzazione delle strutture universitarie, e per difficoltà materiali, almeno al momento, che obstano alla formazione di uno spin-off) assegnare la titolarità del diritto ai coordinatori delle ricerche, regolare con contratti l'opera dei collaboratori, ed affidare a strutture universitarie (corpora.unito.it) la sola distribuzione (accesso web) dei corpora.

3.1 LIBERTÀ DI SFRUTTAMENTO ECONOMICO DEL CORPUS. Lo strumento giuridico attraverso il quale attribuire ai terzi l'esercizio di talune, o tutte, delle prerogative riservate dalla legge sul diritto d'autore al titolare del "corpus" (prerogative derivanti dall'inquadramento giuridico del "corpus" di cui ai precedenti paragrafi (cfr. §§ 1.1-3) sarà rappresentato da un contratto di licenza, concluso tra quest'ultimo ed i singoli utenti-contraenti dal medesimo individuati (e nel rispetto dei principi in materia di contratti traslativi di diritti di utilizzazione ai sensi degli artt. 107 e ss. l.a.).

Il tema della gratuità od onerosità della licenza (o di parte di essa) può anche essere eventualmente rimeditato, dato il desiderio espresso dal gruppo di generare fondi dalla "messa a disposizione" del "corpus" al fine di finanziare i lavori.

3.2 UTILIZZO DELLE LICENZE "CREATIVE COMMONS". Nel progetto di creazione e sfruttamento del "corpus linguistico", le licenze "Creative Commons" possono essere utilizzate tra i soggetti giuridici, e con il contenuto, qui di seguito individuati per sommi capi.

(1) Per l'acquisizione dai singoli autori o aventi causa delle opere originarie le cui elaborazioni formeranno il corpus si deve utilizzare una licenza che, almeno, consenta di (a) tokenizzare, markuppare e taggare il testo, (b) inserirlo in uno o più corpus linguistici, (c) consentire le operazioni di estrazione del testo dai corpora, vietando però espressamente qualsiasi ulteriore attività di ripubblicazione del testo estratto dal corpus medesimo. Un testo rilasciato sotto licenza CC "Attribution" può essere liberamente trattato ed utilizzato in un corpus.

(2) I singoli elaboratori e colui che dirige ed organizza il "corpus" (se vi è un soggetto preposto a tale attività) avranno (o pattuiranno) con l'Università (o altro soggetto individuato quale coordinatore dell'opera) accordi contrattuali (si può utilizzare la licenza CC "Attribution". In questo modo i singoli testi tokenizzati possono essere liberamente utilizzati da terzi col fine di formare nuovi Corpora.

(3) Il “Corpus” così creato è attribuito in capo all’Università (o ad altro soggetto che funge da coordinatore dell’opera collettiva “Corpus” così creata) cui spettano i diritti di sfruttamento che potranno formare oggetto di atto di disposizione a favore di terzi-utenti legittimi tramite licenza CC “Attribution” o CC “Attribution-ShareAlike”. Il Corpus, infatti, costituente opera collettiva e quindi oggetto creativo autonomo rispetto ai singoli testi tokenizzati, è licenziabile in maniera diversa rispetto ai singoli testi non tokenizzati.

4. APPROFONDIMENTI LEGALI. Alla luce di quanto sopra, si ritiene di poter approfondire qui di séguito gli istituti giuridici richiamati nei punti che precedono.

4.1 LA DOPPIA TUTELA GIURIDICA DELLA BANCA DI DATI. La banca di dati, secondo il dettato normativo (Art. 1 n. 9, l.a.), è intesa come un’opera di compilazione ed esattamente una «raccolta di opere, dati o altri elementi indipendenti sistematicamente o metodicamente disposti ed individualmente accessibili mediante mezzi elettronici o in altro modo», ed è opera dell’ingegno protetta oggi dalla legge sul diritto d’autore in base a quanto disposto dagli artt. 64 *quinquies* e *sexies* (Sezione VII “Banche dati”), e dagli artt. 102 *bis* e *ter* (Titolo II bis “Disposizioni sui diritti del costitutore di una banca di dati diritti ed obblighi dell’utente”) della l.a.[6].

La tutela apprestata da tali disposizioni attribuisce al titolare, alle condizioni ivi stabilite, due diritti a sé stanti: il diritto d’autore e il diritto *sui generis*[7] che qui di seguito verranno esaminati.

4.2 IL DIRITTO D’AUTORE (ARTT. 64 QUINQUIES E SEXIES). Sono tutelate dalle disposizioni di cui agli artt. 64 *quinquies e sexies* le banche di dati che siano dotate di “creatività” e siano quindi espressione di una personale concezione o arbitrio valutativo dell’autore. La “creatività” (la capacità di dotare l’opera di un certo grado di personalità propria dell’autore) può attenere, in via alternativa o cumulativa, (j) alla scelta dei materiali da incorporare nella banca di dati e (ij) alla loro modalità di disposizione.

È escluso dalla tutela *de quo* (a meno che non sia di per sé “creativo”) il “contenuto” informativo della banca di dati nella sua interezza, in quanto la protezione ad essa accordata attiene esclusivamente alle modalità “creative” di scelta e disposizione del materiale. Il contenuto dell’opera potrà, se del caso, trovare tutela in forza del diverso diritto *sui generis* attribuito dall’art. 102 *ter*, nella misura in cui «il conseguimento, la verifica e la presentazione di tale contenuto» abbia richiesto un «investimento rilevante» (sul punto, cfr. *infra* § 4).

L’art. 64 *quinquies* a l. a. attribuisce all’autore di una banca di dati taluni diritti di esclusiva:

---

[6] Qualora, tuttavia, l’opera compilativa non possa essere qualificata come banca di dati, in quanto la raccolta non sia sistematicamente o metodicamente disposta, ovvero composta di elementi individualmente accessibili (ma non pare essere questo il caso dei Corpora) ed a condizione che comunque essa sia dotata di un *quid* di “creatività” ad essa dovrà riconoscersi la tutela propria delle opere dell’ingegno, per es. quella predisposta per le opere collettive, ma non sarà ad essa applicabile la specifica disciplina prevista per le banche dati di cui agli artt. 64 *quinquies* e *sexies*. Infatti, anche prima dell’attuazione della direttiva sulle banche di dati, la raccolta di opere era ugualmente suscettibile di protezione in base alle norme sul diritto d’autore, come opera collettiva definita dall’art. 3 l.a. come quelle opere «costituite dalla riunione di opere o di parti di opere, che hanno carattere di creazione autonoma come risultato della scelta e del coordinamento ad un determinato fine letterario, scientifico, didattico, religioso, politico od artistico, quali le enciclopedie, i dizionari, le antologie, le riviste e i giornali».

[7] Si ritiene opportuno segnalare che le banche di dati non tutelabili dal diritto d’autore, in quanto carenti di creatività nella scelta e disposizione dei materiali, possono essere accedere alla protezione accordata dal diritto “*sui generis*” di cui all’art. 102 *bis* e *ter* . La nozione di banca di dati è equivalente sia ai fini del riconoscimento del diritto d’autore che del diritto *sui generis*.

«il diritto esclusivo di eseguire o autorizzare: (a) la riproduzione permanente o temporanea, totale o parziale, con qualsiasi mezzo e in qualsiasi forma; (b) la traduzione, l'adattamento, una diversa disposizione e ogni altra modifica; (c) qualsiasi forma di distribuzione al pubblico dell'originale o di copie della banca dati; (d) qualsiasi presentazione, dimostrazione o comunicazione in pubblico, ivi compresa la trasmissione effettuata con qualsiasi mezzo ed in qualsiasi forma, (e) qualsiasi riproduzione, distribuzione, nonché qualsiasi riproduzione distribuzione, comunicazione, presentazione o dimostrazione in pubblico dei risultati delle operazioni di cui alla lettera b)».

Non sono invece soggetti all'autorizzazione di cui all'art. 64 *quinquies* da parte del titolare del titolare del diritto:

«1. (...) (a) l'accesso e la consultazione *della banca di dati quando abbiano esclusivamente* finalità didattiche e di ricerca scientifica*, non svolta nell'ambito di un'impresa, purché si indichi la fonte e nei limiti di quanto giustificato dallo scopo non commerciale perseguito. Nell'ambito di tali attività di accesso e consultazione le eventuali operazioni di riproduzione permanente della totalità o di parte sostanziale del contenuto su altro supporto sono comunque soggette all'autorizzazione del titolare del diritto; (b) l'impiego di una banca di dati per fini di sicurezza pubblica o per effetto di una procedura amministrativa o giurisdizionale.*

2. Non sono soggette all'autorizzazione dell'autore le attività indicate nell'art. 64 quinquies poste in essere da parte dell'utente legittimo della banca dati e per il suo normale impiego; se l'utente legittimo è autorizzato ad utilizzare solo una parte della banca dati, il presente comma si applica unicamente a tale parte».

4.3 IL DIRITTO "SUI GENERIS" (ARTT. 102 BIS E TER). La nuova disciplina ha introdotto a favore del costitutore della banca di dati un diritto *"sui generis"*. Tale diritto ha ad oggetto il contenuto informativo della banca di dati, nel momento in cui «il conseguimento, la verifica e la presentazione di tale contenuto abbia richiesto un investimento rilevante»[8] . Nel silenzio sia della direttiva CE 96/9 che del D. L.vo 169/1999 sul significato da attribuire alla definizione di "investimento rilevante", in dottrina si è osservato come la banca di dati sarebbe frutto di investimento rilevante, ogni qualvolta essa derivi dall'impiego di ingenti risorse economiche (impiegate nella raccolta, ovvero nella elaborazione dei dati), o, alternativamente o cumulativamente, da sforzi intellettuali ed organizzativi, da apprezzare in rapporto al livello medio del settore di riferimento.

Dispone l'art. 102 bis l.a. 3° comma che:

«3. Indipendentemente dalla tutelabilità della banca di dati a norma del diritto d'autore o di altri diritti e senza pregiudizio dei diritti sul contenuto o parti di esso, il costitutore di una banca dati ha il diritto, per la durata ed alle condizioni stabilite dal presente Capo, di vietare le operazioni di estrazione ovvero di reimpiego della totalità o di una parte sostanziale della stessa. [...]

6. Il diritto del costitutore della banca dati sorge al momento del completamento della banca di dati e si estingue trascorsi quindici anni dal 1. gennaio dell'anno successivo alla data del completamento dello stesso[9] [...]

---

[8] Che non consiste in una vera e propria "creatività". Si è autorevolmente osservato come «Il requisito dell'investimento qualitativamente rilevante sembra poter coprire anche quei casi in cui non vi sia una spendita consistente di risorse quantitativamente valutabili, né una vera e propria creatitività nel senso autoristico tradizionale del termine, ma piuttosto un'idea originale, non tutelabile in base al diritto d'autore o altra disciplina propria del diritto industriale».

[9] Dalla lettura della norma si evince che il contenuto del diritto consiste anzitutto nella facoltà spettante al costitutore della banca di dati di (j) vietare l'estrazione o il reimpiego della *totalità o di una parte sostanziale* del contenuto della banca di dati, e di (ij) vietare l'estrazione o il reimpiego *ripetuti e sistematici di parti non sostanziali* del contenuto della banca dati che presuppongano operazioni contrarie alla normale gestione della banca di dati o che arrechino un pregiudizio ingiustificato ai suoi legittimi interessi. La tutela non attribuisce invece il

8. Se vengono apportate al contenuto della banca di dati modifiche o integrazioni sostanziali comportanti nuovi investimenti rilevanti ai sensi del comma 1, lettera a), dal momento del completamento o della prima messa a disposizione del pubblico della banca di dati così modificata o integrata e come tale espressamente identificata, decorre un autonomo termine di durata della protezione pari a quello di cui ai commi 6 e 7.

9. Non sono consentiti l'estrazione o il reimpiego ripetuti e sistematici di parti non sostanziali del contenuto della banca di dati qualora presuppongano operazioni contrarie alla normale gestione della banca di dati o arrechino un pregiudizio ingiustificato al costitutore della banca di dati[10].

10. Il diritto di cui al comma 3 può essere acquistato o trasmesso in tutti i modi e forme consentiti dalla legge.》

Ai fini della presente disposizione, per «costitutore di una banca dati»[11] si intende chi «effettua investimenti per la costituzione di una banca di dati o per la sua verifica o la sua presentazione, impegnando, a tal fine, mezzi finanziari, tempo o lavoro».

All'utente legittimo della banca di dati messa a disposizione del pubblico vengono attribuite una serie di prerogative individuate nell'art. 102 ter:

«*1. L'utente legittimo della banca di dati messa a disposizione del pubblico non può arrecare pregiudizio al titolare del diritto d'autore o di un altro diritto connesso relativo ad opere o prestazioni contenute in tale banca.*

2. L'utente legittimo di una banca dati messa in qualsiasi modo a disposizione del pubblico non può eseguire operazioni che siano in contrasto con la normale gestione della banca di dati o che arrechino un ingiustificato pregiudizio al istitutore della banca di dati.

3. Non sono soggette all'autorizzazione del costitutore della banca di dati messa per qualsiasi motivo a disposizione del pubblico le attività di estrazione o di reimpiego di parti non sostanziali, valutate in termini qualitativi e quantitativi, del contenuto della banca di dati per qualsivoglia fine effettuate dall'utente legittimo. Se l'utente legittimo è autorizzato ad effettuare l'estrazione o il reimpiego solo di una parte della banca di dati, il presente comma si applica unicamente a tale parte.»[12]

4.4 BANCA DI DATI COME OPERA COLLETTIVA. Le banche di dati, qualora siano «costituite dalla riunione di opere o di parti di opere che hanno carattere di creazione autonoma», e dunque tutelate di per sé come opere dell'ingegno[13], saranno qualificabili come "opere collettive", sot-

---

diritto di vietare a terzi la costituzione di una banca dati equivalente, accedendo autonomamente ad altre fonti informative.

[10] Per *"estrazione"* il disposto normativo intende: *"il trasferimento permanente o temporaneo della totalità o di una parte sostanziale del contenuto di una banca di dati su un altro supporto con qualsiasi mezzo o in qualsivoglia forma"*, e per *"reimpiego"*: *"qualsivoglia forma di messa a disposizione del pubblico della totalità o di una parte sostanziale del contenuto della banca di dati mediante distribuzione di copie, noleggio, trasmissione effettuata con qualsivoglia mezzo e in qualsiasi forma"*.

[11] Si è correttamente osservato che il costitutore della banca di dati non necessariamente dovrà essere un imprenditore commerciale.

[12] L'art. 102 *bis* e *ter* non prevede ulteriori forme di utilizzazioni libere, né in capo all'utente legittimo, né in capo ad altri soggetti, nonostante per entrambi tali diritti la direttiva CE 96/9 consentisse agli stati membri di introdurre discrezionalmente alcune specifiche limitazioni del diritto d'autore e del diritto "*sui generis"*. In particolare, all'utente legittimo della banca dati potevano essere consentite, senza autorizzazione, dell'autore o del costitutore, l'estrazione o il reimpiego di una parte sostanziale del contenuto della banca di dati (a) qualora si trattasse di un'*estrazione per fini privati* del contenuto di una banca di dati non elettronica, (b) qualora si trattasse di *un'estrazione per finalità didattiche o di ricerca scientifica* purché l'utente legittimo ne citasse la fonte, ed in quanto ciò fosse giustificato dagli *scopi non commerciali* perseguiti, (c) qualora si trattasse di estrazione o reimpiego per fini di sicurezza pubblica o per una procedura amministrativa (art. 9 della direttiva). Nulla vieta al costitutore della banca di dati di autorizzare tali attività nonostante il silenzio del disposto normativo.

[13] L'opera collettiva può anche consistere nella raccolta o riunione di elementi non costituenti opere autonome che dia luogo ad un'opera avente carattere rappresentativo.

tostando altresì alla disciplina apprestata dalle disposizioni di cui agli artt. 3, 7 e 38 l.a e saranno protette come opere originali «indipendentemente e senza pregiudizio dei diritti d'autore sulle opere o sulle parti di opere di cui sono composte» (art. 3 l.a.). L'opera collettiva è quindi protetta come opera a sé stante, ma senza che ciò possa comprimere i diritti di utilizzazione economica sui singoli componenti l'opera nel suo complesso il cui esercizio resta riservato, salvo diversa pattuizione contrattuale, in capo agli autori dei singoli contributi i quali eserciteranno liberamente tutte quelle prerogative che non attengano all'inserimento della propria opera nell'opera collettiva (quali, in particolare, il diritto ad utilizzazioni separate della singola opera, il diritto di modifica costituente un rifacimento sostanziale della singola opera, quello di trasformazione in altra forma dell'opera originaria, di traduzione, di adattamento, di riduzione ed ogni altra forma di elaborazione e di trasformazione dell'opera originaria e più in generale di tutti i diritti esclusivi attribuiti dalla legge sul diritto d'autore).

I diritti di utilizzazione economica dell'opera collettiva sono attribuiti in capo a chi abbia organizzato e diretto la creazione dell'opera (art. 7 l.a.), mentre il diritto di utilizzare i singoli "contributi" è riservato ai singoli collaboratori, con l'osservanza degli eventuali patti convenuti (art. 38 2° co., l.a.). Tale disposizione è oggi in dottrina intesa nel senso che l'organizzazione e direzione dell'attività creativa di più collaboratori può essere attribuita in capo a chiunque ed indipendentemente dalla qualifica di imprenditore e dunque compresi gli enti di studio e di ricerca senza scopo di lucro.

4.5 TITOLARITÀ DEI DIRITTI DI UTILIZZAZIONE ECONOMICA DELLA BANCA DI DATI. Il nostro ordinamento giuridico, che stabilisce il principio che solo una persona fisica può acquistare a titolo originario il diritto (morale e patrimoniale) su di un'opera dell'ingegno, fa salvo il principio in forza del quale le creazioni di opere effettuate in esecuzione di rapporti contrattuali che legano l'autore ad altro soggetto giuridico possono legittimamente attribuire direttamente in capo a quest'ultimo, i (soli) diritti di utilizzazione economica dell'opera[14].

In ogni caso, il titolo dell'acquisto è sempre il contratto con l'autore e quindi anche per le banche di dati varranno i principi generali in forza dei quali per attribuire i diritti di utilizzazione economica dell'opera ad un soggetto giuridico non persona fisica occorrerà un contratto di lavoro autonomo o subordinato (tra l'autore o gli autori della banca di dati ed il soggetto giuridico in capo al quale si intende attribuire i diritti di utilizzazione dell'opera) che statuisca tale principio[15].

In forza degli artt. 64 *quinquies* e *sexies* e 102 bis i diritti esclusivi sulla banca dati, spettano, rispettivamente, all'autore della banca di dati ed al suo costitutore. Trattandosi, nel caso dei Corpora, di opera collettiva è considerato "autore" chi "dirige ed organizza" l'opera ed è considerato costitutore chi sostiene l'"investimento rilevante". In entrambe le ipotesi, in applicazione dei principi generali appena delineati, il titolo dell'acquisto sarà rappresentato sempre dal contratto (di lavoro autonomo o subordinato) con l'autore dell'opera e si intenderanno trasferiti i soli diritti patrimoniali in esso ricavabili.

Ci si chiede ora se tale soggetto giuridico titolare dei diritti (di autore e sui generis) afferenti la banca di dati possa essere un ente pubblico, ed in particolare, un istituto universitario, come l'Università di Torino. Si ritiene che nulla osti a che un ente pubblico possa essere titolare dei diritti di utilizzazione economica di una banca di dati[16]. Tale assunto è ricavabile sia dall'inter-

---

[14] Diversamente (pur senza approfondire il tema) i diritti morali d'autore sono indisponibili e intrasferibili.

[15] Tale assunto deve intendersi operante anche nell'ipotesi di opera collettiva, quale è la banca di dati, al fine di rendere effettivo il trasferimento dei diritti di utilizzazione economica dell'opera in capo a chi "organizza e dirige" i lavori altrui.

[16] Si ritiene comunque in ogni caso opportuno un esame delle norme statutarie e regolamentari dell'Ente in questione al fine di un'analisi, in concreto, delle eventuali limitazioni al potere contrattuale dell'Ente.

pretazione dell'art. 38 l.a. che, secondo accreditata dottrina, attribuirebbe a qualunque soggetto giuridico (e quindi non soltanto all'impresa editoriale) che assuma l'onere ed il rischio della creazione dell'opera collettiva (inclusi enti di studio e di ricerca senza scopo di lucro) i diritti di utilizzazione economica della medesima. Resta poi salvo il principio espresso dall'art. 11 2° comma l.a in forza del quale agli enti pubblici culturali spetta sempre il diritto d'autore sulla "raccolta dei loro atti e sulle loro pubblicazioni".

BIBLIOGRAFIA.

ALLORA - BARBERA
¶ 5 Adriano Allora - Manuel Barbera, *Il problema legale dei corpora. Prime approssimazioni*, in questo volume, pp. 109-118.

BARBERA - CORINO - ONESTI
¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

CIURCINA - RICOLFI
¶ 7 Marco Ciurcina - Marco Ricolfi, *Le Creative Commons Public Licences per i corpora. Una suite di modelli per la linguistica dei corpora*, in questo volume, pp. 127-132.

HEID
¶ 4 Ulrich Heid, *Il corpus WorkBench come strumento per la linguistica dei corpora. Principi ed applicazioni*, in questo volume, pp. 89-108.

SITI DI RIFERIMENTO.

bmanuel.org	`http://www.bmanuel.org/projects/index.html`
corpora.unito.it	`http://www.corpora.unito.it/`
CC	`http://creativecommons.org/`
GNU	`www.gnu.org` `www.fsf.org`
Linux Facile	`http:www.linuxfacile.org`

MARCO CIURCINA - MARCO RICOLFI
(Università di Torino)

7. Le *Creative Commons Public Licences* per i corpora. *Una suite di modelli per la linguistica dei corpora.*

0. PREMESSA. Il processo che porta alla creazione di un corpus, inteso e definito come in Barbera - Corino - Onesti qui ¶ 3 (cfr. la definizione in § 4) incrocia il diritto d'autore in tre diversi momenti:

- acquisizione dei diritti sulle opere sulle quali si effettua il trattamento computazionale (tokenizzazione, markup ed eventuale tagging);
- acquisizione dei diritti da parte di coloro i quali realizzano il trattamento computazionale delle opere;
- utilizzazione del Corpus da parte di terzi.

Nelle prime due fasi si devono utilizzare modelli di licenza che consentono al coordinatore del Corpus di acquisire tutti i diritti necessari per licenziare al pubblico il Corpus stesso secondo il modello di licenza scelto.

Per la terza fase si propone di utilizzare una *Creative Commons Public Licence*: la "CCPL Attribuzione Condividi allo stesso modo" (*CCPL Attribution Share Alike*). Questa licenza appare infatti idonea a massimizzare la libera circolazione del Corpus stesso e delle opere elaborate mediante trattamento computazionale (cfr. qui Zanni ¶ 6, § 3.1).

0.1 CREATIVE COMMONS PUBLIC LICENSES. Le *Creative Commons Public Licenses* (CCPL) sono 6 modelli di licenza di diritto d'autore realizzate da Creative Commons (CC) con lo scopo di favorire la creazione di contenuti per i quali solo alcuni, ben specificati, diritti sono stati riservati a priori dagli autori; fatti salvi tali diritti, tutti gli altri usi sono esplicitamente consentiti: "alcuni diritti riservati".

Le sei CCPL sono generate dalla combinazione delle seguenti quattro opzioni:

- *Attribuzione*: l'autore dell'opera deve sempre essere indicato;
- *Non commerciale*: l'opera non può essere usata a fini di lucro;
- *Non opere derivate*: non è consentita la creazione di opere derivate (per esempio, la traduzione in un'altra lingua);
- *Condividi allo stesso modo*: eventuali opere derivate devono essere rilasciate sotto la stessa CCPL dell'opera originale.

1	Attribuzione
2	Attribuzione - Non opere derivate
3	Attribuzione - Non commerciale - Non opere derivate
4	Attribuzione - Non commerciale
5	Attribuzione - Non commerciale - Condividi allo stesso modo
6	Attribuzione - Condividi allo stesso modo.

Tav. 1: Le sei licenze standard CC.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp.127-132.

L'opzione Attribuzione è obbligatoria dalla versione 2.0 delle CCPL. Oggi è disponibile la versione 2.5 delle Licenze. La terza e la quarta opzione sono incompatibili tra loro: con la Condividi allo stesso modo (*Share Alike*), infatti, si concede il diritto ai terzi di realizzare opere derivate sotto certe condizioni (quelle scelte dall'autore originario).

Le combinazioni delle 4 opzioni rendono quindi possibili almeno 6 modelli di licenze, compendiati nella Tav. 1.

Utilizzando gli strumenti disponibili nel sito di creativecommons.org si ottengono le istruzioni per associare la licenza appropriata alla propria opera, e la licenza medesima in tre diversi formati:

- il "Commons Deed", un semplice riassunto della licenza, corredato di apposite icone per favorirne la comprensione;
- il "Legal Code", la licenza vera e propria, scritta in linguaggio "legalese";
- il "Digital Code", una traduzione della licenza in codici interpretabili dagli elaboratori elettronici per permettere ai motori di ricerca e ad altre applicazioni di identificare il tipo di licenza associato all'opera.

1. I MODELLI. Seguono tre modelli di licenza, uno per ciascuno dei tre momenti individuati sopra:

a acquisizione dei diritti sulle opere da trattare mediante elaborazione computazionale;
b acquisizione dei diritti sull'elaborazione computazionale delle opere;
c licenza del Corpus.

I modelli scelti sono stati costruiti alla luce dell'obiettivo di massimizzare il riuso (per questo orientamento cfr. qui Barbera ¶ 1, § 2.1.d) delle opere trattate computazionalmente acquisendo i diritti necessari per l'uso di queste da parte dei fornitori.

Pare legittimo ritenere che la scelta di modelli di licenza "aperti a valle" non debba costituire un problema per i fornitori di opere i quali, in ogni caso, sono pienamente tutelati sul piano giuridico, giacché con i modelli di licenza proposti non concedono nessun diritto d'uso delle opere in versione originale[1].

1.1 IL CONTRATTO FORNITORI. Questo modello di licenza si applica all'acquisizione dei diritti sulle opere da trattare mediante elaborazione computazionale (e da inserire nel corpus) dai titolari di questi.

Il contratto fornitori, fermo il divieto di ripubblicare l'opera nel formato originale, consente al coordinatore del corpus di utilizzare l'opera per:

- fare trattamento computazionale delle opere licenziate;
- inserire le opere trattate mediante elaborazione computazionale, o parte di esse, in uno o più Corpora linguistici ed estrarle da questi nella loro versione trattata mediante elaborazione computazionale;
- utilizzare le opere estratte da un Corpus nella loro versione trattata mediante elaborazione computazionale, o parte di esse, per qualsiasi scopo.

Il modello per l'acquisizione di materiali da trasformare in corpora è pertanto il seguente:

[1] Giusta l'impostazione proposta in Allora - Barbera ¶ 5, § 3.1; è comunque possibile utilizzare modelli di licenza diversi per far fronte a necessità specifiche di ricerca limitando, per esempio, il nòvero di usi consentiti delle opere trattate computazionalmente ed intervenendo simmetricamente sui contratti con i fornitori e collaboratori e sulla licenza del corpus.

DICHIARAZIONE

Il sottoscritto ______________________________, nato a ______________ il __________ e residente in ____________ via ______________________ n.____, C.F. ______________, C.I. n. ______________ rilasciata dal Comune di ______________ in data __________ (di seguito "Licenziante"), legittimo titolare di tutti i diritti di utilizzazione economica qui trasferiti, come dichiara e garantisce,

CONSENTE

a ____________________ (di seguito "Licenziatario"), ed a chiunque sia da esso autorizzato, di utilizzare

- ______________________
- ______________________
- ______________________

(di seguito "Opere" ed "Opera" con riferimento a ciascuna di esse) col fine di svolgere su tali Opere attività di:

- trattamento computazionale consistente nella tokenizzazione e markuppatura e nell'eventuale tagging del testo delle Opere (di seguito "Trattamento Computazionale");
- inserimento delle Opere modificate mediante Trattamento Computazionale (di seguito "Opere Trattate"), o di parte di esse, in uno o più Corpora linguistici;
- estrazione delle Opere Trattate, o di parte di esse, da uno o più Corpora linguistici;
- uso delle Opere Trattate estratte da un Corpus, o di parte di esse, per qualsiasi scopo.

Il Licenziante concede espressamente al Licenziatario ed ai suoi aventi causa facoltà di:

- riprodurre, distribuire, comunicare, presentare o dimostrare in pubblico, in qualsiasi modo o forma, le Opere Trattate, o parte di esse, ove tali facoltà siano necessarie per realizzare le attività di cui sopra (Trattamento Computazionale, inserimento in, ed estrazione da Corpora, uso delle Opere Trattate estratte da Corpora);
- licenziare i Corpora nei quali siano inserite le Opere Trattate secondo i termini d'una licenza Creative Commons Public Licence Attribuzione Condividi allo stesso modo.

La licenza non comprende il diritto di riprodurre, distribuire, comunicare al pubblico, presentare o dimostrare in pubblico l'Opera nella sua versione originaria, e cioè rimuovendo od occultando le modifiche realizzate mediante Trattamento Computazionale, senza l'espressa autorizzazione scritta del Licenziante.

La licenza è concessa a titolo gratuito e per tutta la durata dei diritti di utilizzazione economica sulle Opere oggetto di licenza[2].

Il Licenziatario si impegna a indicare in ogni riproduzione delle Opere all'interno di un Corpus la seguente dicitura: «© *[nome del titolare dei diritti di utilizzazione economica], [anno di pubblicazione] – Diritti sull'opera nella sua versione originaria riservati.* L'assolvimento di tale onere potrà avvenire anche in un'unica soluzione con riferimento ad una pluralità di Opere, e comunque con modalità tali che il riferimento al titolare dei diritti di utilizzazione economica risulti in modo non equivoco. A tal fine il Licenziante si obbliga a fornire al Licenziatario tutte le necessarie indicazioni.

[Luogo], li ________

[Firma]

Tav. 2: Il modello del contratto-fornitori.

[2] Nel caso in cui il fornitore conceda licenza dietro pagamento di un corrispettivo si sostituisca questa frase con: "La licenza è concessa a fronte del pagamento dell'importo di __________ e per tutta la durata dei diritti di utilizzazione economica sulle Opere oggetto di licenza."

1.2 IL CONTRATTO COLLABORATORI. Se il modello di licenza precedente (§ 1.1, Tav. 2) si applica all'acquisizione dei diritti da parte dei collaboratori che realizzano il trattamento computazionale delle opere da inserire nel Corpus, il "contratto collaboratori" consente al coordinatore del Corpus di utilizzare le elaborazioni computazionali realizzate dal collaboratore per:

- fare ulteriore trattamento computazionale;
- inserire le opere trattate mediante elaborazione computazionale, o parte di esse, in uno o più Corpora ed estrarle da questi nella loro versione trattata computazionalmente;
- utilizzare le opere estratte da un Corpus nella loro versione trattata mediante elaborazione computazionale, o parte di esse, per qualsiasi scopo.

DICHIARAZIONE

Il sottoscritto ______________________________, nato a ______________ il __________ e residente in ____________ via ________________ n.____, C.F. ______________, C.I. n. ______________ rilasciata dal Comune di ______________ in data __________ (di seguito "Licenziante"), legittimo titolare di tutti i diritti di utilizzazione economica qui trasferiti, come dichiara e garantisce, nel quadro della propria attività di collaborazione alla realizzazione del Corpus linguistico ________ [indicare Corpus], mediante realizzazione di attività di tokenizzazione, markuppatura ed eventuale tagging di testi (di seguito "Trattamento Computazionale")

CONSENTE

a ____________ (di seguito "Licenziatario"), ed a chiunque sia da esso autorizzato, di utilizzare il Trattamento Computazionale da esso realizzato sulle seguenti opere:

- ______________________
- ______________________
- ______________________

(di seguito "Opere" ed "Opera" con riferimento a ciascuna di esse) col fine di svolgere su tali Opere modificate mediante Trattamento Computazionale (di sèguito "Opere Trattate") attività di:

- ulteriore Trattamento Computazionale;
- inserimento delle Opere Trattate, o di parte di esse, in uno o più Corpora linguistici;
- estrazione delle Opere Trattate, o di parte di esse, da uno o più Corpora linguistici;
- uso delle Opere Trattate estratte da un Corpus, o di parte di esse, per qualsiasi scopo.

Il Licenziante concede espressamente al Licenziatario ed ai suoi aventi causa facoltà di:

- riprodurre, distribuire, comunicare, presentare o dimostrare in pubblico, in qualsiasi modo o forma, le Opere Trattate, o parte di esse, ove tali facoltà siano necessarie per realizzare le attività di cui sopra (Trattamento Computazionale, inserimento in, ed estrazione da Corpora, uso delle Opere Trattate estratte da Corpora);
- licenziare i Corpora nei quali siano inserite le Opere Trattate secondo i termini d'una licenza Creative Commons Public Licence Attribuzione Condividi allo stesso modo.

La licenza è concessa a titolo gratuito e per tutta la durata dei diritti di utilizzazione economica sulle Opere Trattate oggetto di licenza.

Il Licenziatario si impegna ad indicare il Licenziante tra i collaboratori del Corpus.

[Luogo], li ________

[Firma]

Tav. 3: Il modello del contratto-collaboratori.

Il modello-tipo per l'acquisizione di materiali da trasformare in corpora è pertanto quello offerto in Tav. 2[3].

1.3 IL CONTRATTO UTILIZZATORI (LA LICENZA CCPL DEI CORPORA). Questo modello di licenza regola il rilascio al pubblico del Corpus realizzato utilizzando le opere acquisite dagli aventi diritto e le elaborazioni computazionali realizzate dai collaboratori.

Il contratto utilizzatori, fermo il divieto di ripubblicare le opere contenute nel Corpus nella loro versione originaria, consente a chiunque di utilizzare il Corpus secondo i termini della licenza CCPL scelta, e cioè la "CCPL Attribuzione-Condividi allo stesso modo, v. 2.5 Italia".

In base a questa licenza chiunque può utilizzare il Corpus nei seguenti modi:

- riprodurre, distribuire, comunicare al pubblico, esporre in pubblico il Corpus,
- creare opere derivate dal Corpus,
- usare il Corpus a fini commerciali,

alle seguenti condizioni:

- **Attribuzione:** si deve riconoscere il contributo dell'autore originario del Corpus;
- **Condividi allo stesso modo:** se si altera, trasforma o sviluppa il Corpus, si può distribuire l'opera risultante solo per mezzo di una licenza identica a questa.

In occasione di ogni atto di riutilizzazione o distribuzione, si devono chiarire agli utenti i termini della licenza del Corpus.

Oltre a quanto sopra è anche concesso il diritto di utilizzare le singole opere contenute nel Corpus, o parte di esse (ma solo nella versione modificata mediante elaborazione computazionale), per:

- farne ulteriore trattamento computazionale;
- inserirle in uno o più Corpora linguistici ed estrarle da questi nella loro versione trattata mediante elaborazione computazionale;
- utilizzarle, ma solo nella loro versione trattata mediante elaborazione computazionale, per qualsiasi scopo.

Il modello di licenza per i corpora è pertanto il seguente:

[3] Nel caso in cui il collaboratore conceda licenza dietro pagamento di un corrispettivo si sostituisca il paragrafo "La licenza è concessa a titolo gratuito e per tutta la durata dei diritti di utilizzazione economica sulle Opere Trattate oggetto di licenza" con il seguente:

> "La licenza è concessa a fronte del pagamento dell'importo di ___________ e per tutta la durata dei diritti di utilizzazione economica sulle Opere oggetto di licenza."

Si segnala però che, nel caso di prestazione a pagamento d'attività di collaborazione all'elaborazione computazionale di opere è necessario coordinare la licenza del collaboratore col rapporto contrattuale volto a regolare gli altri aspetti del rapporto.

LICENZA

Questo Corpus è utilizzabile secondo i termini della licenza Creative Commons Attribuzione-Condividi allo stesso modo 2.5 Italia:

`http://creativecommons.org/licenses/by-sa/2.5/it/`

La licenza Creative Commons Attribuzione-Condividi allo stesso modo 2.5 Italia non si applica alle opere a sé stanti contenute nel Corpus che pertanto restano soggette ai termini di licenza indicati per ciascuna di esse.

Le singole opere contenute nel Corpus possono essere utilizzate nei modi e termini di seguito specificati esclusivamente nella versione modificata mediante tokenizzazione, markuppatura ed eventuale tagging (di seguito "Trattamento Computazionale"). È pertanto espressamente escluso qualsiasi diritto di riprodurre, distribuire, comunicare al pubblico, presentare o dimostrare in pubblico le opere contenute nel Corpus nella loro versione originaria, e cioè rimuovendo od occultando le modifiche realizzate mediante Trattamento Computazionale.

Le singole opere contenute nel Corpus e modificate mediante Trattamento Computazionale (di seguito "Opere Trattate") possono essere utilizzate col fine di svolgere su tali Opere Trattate attività di:

- ulteriore Trattamento Computazionale;
- inserimento delle Opere Trattate, o di parte di esse, in uno o più Corpora linguistici;
- estrazione delle Opere Trattate, o di parte di esse, da uno o più Corpora linguistici;
- uso delle Opere Trattate estratte da un Corpus, o di parte di esse, per qualsiasi scopo.

È concessa facoltà di riprodurre, distribuire, comunicare, presentare o dimostrare in pubblico, in qualsiasi modo o forma, le Opere Trattate, o parte di esse, ove tali facoltà siano necessarie per realizzare le attività di cui sopra (Trattamento Computazionale, inserimento in, ed estrazione da Corpora, uso delle Opere Trattate estratte da Corpora).

Tav. 3: Il modello del contratto-utilizzatori ("Licenza").

BIBLIOGRAFIA.

ALLORA - BARBERA
¶ 5 Adriano Allora - Manuel Barbera, *Il problema legale dei corpora. Prime approssimazioni*, in questo volume, pp. 109-118.

BARBERA
¶ 1 Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in questo volume, pp. 3-20.

BARBERA - CORINO - ONESTI
¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

SITI DI RIFERIMENTO.

bmanuel.org	`http://www.bmanuel.org/projects/index.html`
corpora.unito.it	`http://www.corpora.unito.it/`
CC	`http://creativecommons.org/`
CC-it	`http://it.creativecommons.org/`

PARTE
.II.

MANUEL BARBERA
(Università di Torino)

8. Un tagset per il Corpus Taurinense[1]. *Italiano antico e linguistica dei corpora.*

Wer Perlen will
der muss ins Meer sich stürzen.
Johann Wolfgang Goethe, Nachlaßstuck zu *West-östlichem Divan*.

0. PREMESSA. In principio (come già si era detto in Barbera ¶ 1, in questo volume) fu *ItalAnt*, ossia il progetto fondato da Lorenzo Renzi e Giampaolo Salvi di una grammatica (o meglio, una sintassi) dell'italiano antico (*lege*: fiorentino duecentesco), ad ideale continuazione della *Grande grammatica* (Renzi - Salvi et alii 1988-1995), basata su un delimitato canone di testi accessibili anche in formato elettronico, che è poi un sottoinsieme della base testuale dell'OVI generosamente messo a disposizione da Pietro Beltrami. Il *Padua Corpus* o *Corpus ItalAnt*, come è di solito informalmente chiamato questo insieme di testi[2], era stato selezionato da Renzi e Salvi (cfr. Renzi 1998, 29) in modo da essere variegato dal punto di vista dei generi testuali rappresentati (lirico, didattico, narrativo, documentario ecc.) ma unitario dal punto di vista diacronico (1250-1300) e diacorico (solo fiorentino), in modo da avvicinarsi il più possibile ad un ideale spaccato sincronico[3], ed era consultabile in ambiente PC con GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), un sistema di ricerca pensato dal suo creatore Domenico Iorio-Fili e dal suo ispiratore Pietro Beltrami per esigenze prevalentemente lessicografiche[4]. Se, però, le finalità del gruppo padovano erano la produzione di una grammatica (per la quale il Padua Corpus era già uno strumento utile), quelle del gruppo torinese di Manuel Barbera e Carla Marello erano semmai di produrre un corpus che si ponesse a pieni titoli nel panorama dell'attuale linguistica dei corpora annotati; e per questa specifica finalità i limiti linguistico-computazionali del Padua Corpus (che, appunto, *non è* un corpus) ci apparvero presto evidenti (cfr. Barbera - Marello 1999/2001, §§ 3 e 5). Così, il risultato del nostro lavoro fu il *Corpus Taurinense*[5] (CT), che è la reincarnazione in un corpus, annotato, tokenizzato[6] ed

---

[1] Il presente contributo è una versione modificata, ampliata ed aggiornata di *Italiano antico e linguistica dei corpora: un tagset per ItalAnt*, relazione presentata al VI Convegno Internazionale SILFI *Tradizione & Innovazione: la linguistica e filologia italiana alle soglie di un nuovo millenio*, Gerhard-Mercator-Universität Duisburg, 28 giugno - 2 luglio 2000, la stampa dei cui *Atti* non è ancora conclusa. L'aggiornamento, si badi però, ha tenuto conto soprattutto dell'attuale stato dei lavori del CT, ma non è stato portato sistematicamente a fondo per quel che riguarda la bibliografia in materia.
[2] Propriamente, infatti, secondo i criteri qui definiti in Barbera - Corino - Onesti ¶ 3, questa raccolta non si qualifica strettamente come "corpus" in senso tecnico, a causa della mancanza di una vera tokenizzazione e per altri minori "difetti" (per cui cfr. appunto Barbera - Marello 1999/2001: §§ 3 e 5).
[3] Per i criteri alla base della selezione del Padua Corpu*s* cfr. Renzi 1998, p. 29; per una loro discussione critica cfr. Barbera - Marello 1999\2000, § 1.
[4] È infatti nato per la gestione della base testuale che è alla base del *Vocabolario Storico della Lingua Italiana* (Beltrami 1983-...) in corso di realizzazione presso l'OVI (Opera del Vocabolario Italiano). Per una presentazione di GATTO cfr. Iorio-Fili 1997.
[5] Il suo nome, analogamente al Padua Corpus, è tratto dalla sede del gruppo cofinanziato.
[6] Per il concetto di tokenizzazione cfr. qui Barbera - Corino - Onesti ¶ 3, §§ 1 ed 1.3. In generale, per la terminologia assai poco puristica cfr. quanto abbiamo argomentato in Barbera - Corino - Onesti ¶ 3 e soprattutto in Barbera - Marello 2003 *i.s.* Ci conforta di essere quanto a ciò in allegra e rispettabil barca. I limiti di ogni purismo, infatti, erano già stati lucidamente evidenziati dal Leopardi, che trovandosi in un simile impaccio, argo-

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp.135-168.

interrogabile tramite il CWB (Corpus Work Bench; cfr. Christ - Schulze 1996) dell'IMS Stuttgart, dei testi ("Padua Corpus") scelti come base per ItalAnt.

Per ottenere questo risultato, ossia un corpus annotato morfosintatticamente secondo i più recenti standard, in modo da renderlo così confrontabile con i corpora esistenti nelle maggiori lingue contemporanee, si sono rese necessarie varie operazioni, spesso complesse e, per così dire, tutte "da inventare": la *corpus annotation*, infatti, è una branca della linguistica computazionale che finora si è occupata solo raramente di corpora "antichi", sicché avevamo pochi precedenti su cui basarci[7]. Non è qui luogo per diffonderci su tutte queste vicende (vi ritorneremo in altra sede); basti ricordare che bisognava tener conto delle specificità dell'italiano antico in relazione tanto agli automatismi computazionali quanto alle esigenze dell'analisi linguistica.

In questo contributo ci soffermeremo invece sul solo aspetto della costruzione del tagset per il POS-tagging[8]. In particolare, discuteremo prima diffusamente delle problematiche sottese alla proposizione di un tagset (cfr. §§ 1-2 e sottoparagrafi), illustreremo funzionamento e struttura delle "gerarchie tipate" (cfr. § 3 e sottoparagrafi), presenteremo quindi il nostro tagset (cfr. § 5 e sottoparagrafi) con poche ulteriori osservazioni[9] (cfr. § 4) e concluderemo dando la "feature declaration" (cfr. § 6 e sottoparagrafi), e producendo un piccolo esempio annotato (cfr. § 7). Per un confronto (in vista di una riunificazione, cfr. *supra* Barbera ¶ 1 § 3.1) tra i vari tagset implementati su bmanuel.org / corpora.unito.it, e per un perfezionamento pratico dei criteri (specie per le *labels*), cfr. *infra* Barbera ¶ 23.

1. I REQUISITI DI UN TAGSET. Le considerazioni che stanno, in generale, alla base della creazione di un tagset e che, di fatto, ci hanno guidato nella elaborazione di questo specifico tagset, sono di natura abbastanza eterogenea. Spesso queste sono lasciate implicite, ma vista la rilevanza pratica e teorica che hanno, sarà forse il caso finalmente di presentarle e discuterle in modo esplicito.

---

mentava nello *Zibaldone* (p. 3195) che «se vuol dunque l'Italia avere una filosofia ed una letteratura moderna filosofica, le quali finora non ebbe mai, le conviene di fuori pigliarle, non crearle da se [*sic*]; [...] e volendole ricevere, nol potrà altrimenti che ricevendo altresì assai parole e frasi di là, ad esse intimamente e indivisibilmente spettanti e fatte proprie» (ed. Pacella 1991, p. 1677; per una citazione più estesa di questo passo, cfr. qui Barbera ¶ iiij). E, *mutatis mutandis*, quasi tutte le osservazioni consegnate alle pagine 3192-3196 di quel grande non hanno affatto perso il loro valore ed attualità. In assenza di buoni traducenti nativi, all'epoca dell'originario contributo per la SILFI (2000: ben sette anni fa) avevamo preferito la cautela, mantenendo in inglese (e pertanto in corsivo, e con plurali in *-s*) quanto diversamente non avremmo bene saputo come chiamare; ora, sentendoci un poco più forti, abbiamo risolutamente adottato la soluzione del prestito non adattato per le forme base (*e.g.* token, pertanto, invariabile ed in tondo) e normalmente affissato per le derivate (*e.g.* tokenizzato, con conservazione grafica nel radicale ma poi suffissazione regolare italiana).

[7] Perdipiù il *Penn-Helsinki Parsed Corpus of Middle English* (PPCME) ed il *Tycho Brahe Parsed Corpus of Historical Portuguese* (TBPCHP), che erano le esperienze più note in questo settore, sono entrambi dei treebank, cioè dei corpora con annotazione puramente sintattica, e presentano pertanto problematiche spesso diverse dalle nostre. Eravamo a conoscenza di alcuni esperimenti di annotazione morfologica presso il CiBIT (Centro interuniversitario Biblioteca Italiana Telematica) di Pisa, ma i loro risultati (sostanzialmente le *Opere di Dante lemmatizzate con marcatori grammaticali* di Mirko Tavoni) sono stati diffusi solo recentemente, ed hanno comunque caratteristiche diverse; dell'esistenza di un *Analizzatore Morfosintattico dell'Italiano Antico* (AMIA, di Fabrizio Beggiato) si è avuta notizia solo dal 2003, né più se ne è saputo alcunché, e, ad ogni buon conto, anche questo progetto avrebbe caratteristiche assai diverse dal nostro (i suoi risultati, ad es., non sarebbero disambiguati). Molto interessanti, invece, i risultati ottenuti da Achim Stein (cfr. la sua homepage e quella del TreeTagger) per l'antico francese, ma anche questi sono stati diffusi solo a partire dal 2003.

[8] Ossia, per il tagging morfosintattico (POS è il normale acronimo per *Part Of Speech*): per il concetto di tagging cfr. Barbera . Corino - Onesti ¶ 3, §§ 1 ed 1.4.

[9] La base dei §§ 4 e 5 è proprio il materiale che avevamo messo fin da sùbito a disposizione dei nostri annotatori: ed è solo a partire dalle loro "reazioni" e dalle nostre riflessioni su cosa incontravano, che è stato possibile arrivare alla versione finale qui presentata.

1.1 CONSENSUALITÀ E NEUTRALITÀ. Una prima istanza **[1]**, quella della "consensualità e neutralità" del sistema di annotazione, è affatto preliminare, e va affrontata sùbito. È stato più volte sottolineato che «it is a good idea for annotation schemes to be based as far as possible on consensual or theory-neutral analyses of the data» (Leech 1997, p. 7). Tale argomento è di natura evidentemente pratica ma ha implicazioni teoriche di non poco momento.

Da un lato, infatti, il requisito di "consensualità" invocato dai linguisti computazionali allo scopo di garantire la massima accessibilità e (ri)utilizzabilità delle loro annotazioni si può facilmente riportare alla nozione di "concetto ingenuo" elaborata da Giorgio Graffi (cfr. Graffi 1991). Dall'altro quello di "neutralità" va inteso propriamente anche come "neutralità metalinguistica": i modelli in cui sono espressi i dati in *corpus linguistics* sono puramente dei metalinguaggi descrittivi e come tali convenzionali[10] che né ambiscono né devono ambire ad identificarsi con le strutture dell'oggetto che descrivono. Non hanno pertanto le stesse caratteristiche epistemologiche e, per così dire, "ontologiche" di teorie linguistiche "forti" come la grammatica generativa, ma non ne sono affatto, di per sé, incompatibili.

È, d'altra parte, in questo ordine di idee che si sono sviluppate le grammatiche *lato sensu* "categoriali" e "ad unificazione" che stanno riportando significativi successi in applicazioni di NLP ("Natural Language Processing") e di *corpus linguistics*[11].

1.2 ADEGUATEZZA DESCRITTIVA E STANDARDIZZAZIONE. Ciò premesso, i due successivi requisiti cui dovrebbe rispondere un tagset possono apparire tra loro in parte contraddittori: [2] "adeguatezza descrittiva" specifica e [3] "standardizzazione" del formato.

Il requisito **[2]** comporta che il modello descrittivo adottato sia il più possibile adeguato a rendere conto della specificità del corpus oggetto. Ad esempio, nel caso dell'italiano antico, abbiamo dovuto introdurre la POS "postposizione" per rendere conto dei vari *meco*, *teco*, *seco* laddove al moderno italiano parlato sarebbe bastata quella di "preposizione".

Il requisito **[3]**, invece, punta in direzione della standardizzazione, ossia della omogeneità e compatibilità con altre esperienze di annotazione di corpora. I vantaggi di ciò sono evidenti: si va dalla riutilizzabilità dei corpora così preparati per ricerche diverse da quella per la quale sono stati costruiti (il passaggio dall'OVI – con finalità lessicografiche – ad ItalAnt – con finalità di descrizione linguistica – ne è già un esempio), alla possibilità di dialogo e scambio di dati tra progetti diversi, cumulando così informazioni estratte da più corpora, alla massima compatibilità con sistemi informatici diversi. L'esigenza che «resources should be reusable, interchangeable, shareable» (Monachini - Calzolari 1999, p. 149) è ormai molto avvertita anche a livello istituzionale: non a caso negli ultimi anni si sono moltiplicate le iniziative internazionali in questo senso (cfr. Monachini - Calzolari 1999, pp. 149-150). Nel nostro caso, poi, la volontà di rendere il CT compatibile e "dialogabile" con gli altri corpora annotati esistenti è particolarmente sentita, data la natura sperimentale ed innovativa della nostra impresa, che speriamo si possa porre un poco come progetto pilota per ulteriori iniziative.

Un ottimo bilanciamento tra le due esigenze sopra denunciate è stato raggiunto, in sede europea, dall'iniziativa EAGLES[12] (Expert Advisory Group on Language Engineering Standards), culminata – per quel che qui ci concerne – nella elaborazione di una serie di *Guidelines*

---

[10] Naturalmente "convenzionale" non è da intendersi come 'arbitrario' ma, come usuale in logica, nel senso del principio di tolleranza di Carnap (cfr. Carnap 1937/1934, pp. 51-52 e 1974/1963, p. 19).

[11] Orientamenti di questo tipo si hanno dalla *Lexical Functional Grammar* ("LFG"; cfr. Kaplan - Bresnan 1982), alla *Head-Driven Phrase Structure Grammar* ("HPSG"; cfr. Pollard - Sag 1987), alla *Constraint Grammar* ("EnCG") sviluppata a partire dal 1990 ad Helsinki per l'inglese (Karlsson et alii 1995; cfr. la homepage di CG2), al *Comprehensive Unification Formalism* ("CUF") sviluppato a Stuttgart (Dörre - Dorna 1993; cfr. la homepage del CUF) ed alle *Categorial Grammars* ("CG") in genere (cfr. König 1996). Per una trattazione recente ed accessibile di questo tipo di grammatiche cfr. Allegranza - Mazzini 2000.

[12] Ora proseguita da ISLE (International Standards for Language Engineering).

o “raccomandazioni” per la annotazione linguistica[13]. La soluzione, in questo caso, sta nel fatto che, una volta accettata una comune struttura formale – quella basata sulla nozione di gerarchia tipata[14] –, si introduce poi una elevata parametricità di dettaglio, distinguendo tra elementi obbligatori e facoltativi. Monachini - Calzolari 1996, in particolare, sia pure sviluppato espressamente per l’annotazione di lessici anziché di corpora, è in questo senso un documento fondamentale, in quanto presenta un accurato confronto tra i più importanti tagset esistenti per le lingue europee, ricavandone le “raccomandazioni” di standardizzazione EAGLES. Il tagset del *Corpus Taurinense* è pienamente conforme a queste *Guidelines* e potrà così dialogare con ogni iniziativa a livello europeo, affiancandosi, ad esempio, alle proposte per l’italiano moderno (Monachini 1996, di solito riferite come “ELM-IT”[15]), per il tedesco (Teufel - Stöckert 1996, cioè “ELM-DE”[16]), per il francese (Rekowski 1995, “ELM-FR”) e per l’inglese (Teufel 1996, “ELM-EN”).

1.3 PRATICITÀ COMPUTAZIONALE. L’ultimo principale requisito di cui tener conto è **[4]** la “praticità computazionale”, cioè la possibilità di gestire computazionalmente un’applicazione, che si riflette poi nell’efficienza di interrogazione e nella disponibilità a generare nuova informazione.

Inevitabilmente, si devono accettare alcune limitazioni tecniche, che, per quanto appaiano “costose” in termini linguistici, si possono a volte tradurre, se accettate consapevolmente e gestite in modo intelligente, in rilevanti vantaggi.

Un esempio è quello del contenimento del tagset. «The POS tagsets used to annotate large corpora in the past have traditionally been fairly extensive. The pioneering Brown Corpus distinguishes 87 simple tags [...] the Lancaster-Oslo/Bergen (LOB) Corpus uses about 135 tags, the Lancaster UCREL group around 165 tags, and the London-Lund Corpus of Spoken English 197 tags[17]» riassumevano Marcus - Santorini - Marcinkiewicz 1994, p. 274, poi argomentando che «however, the stochastic orientation of the Penn Treebank and the resulting concern with sparse data led us to modify the Brown Corpus tagset by pairing it down considerably»[18]. La contrapposizione, in effetti, è tra grandi tagset[19] applicati manualmente o (semi) automaticamente tramite grammatiche di microregole[20] (e nessuno di questi, inoltre, è costruito per gerarchie tipate) e tagset pensati per essere applicabili da un tagger stocastico. Se, poi, si limita il tagset a non più di 70 tag[21] gerarchici, il corpus così annotato avrà un rendimento ottimale come *training corpus* per un annotatore stocastico (cfr. Heid 1998).

[13] Cfr. Leech - Wilson 1999 e Monachini - Calzolari 1999.

[14] Cioè su feature gerarchiche con ereditarietà: ne parleremo più diffusamente tra poco.

[15] Delle analoghe e stimolanti esperienze condotte da Marco Baroni e dalla sua equipe (cfr. Baroni et alii 2004) non potevamo ovviamente tener conto per ovvie ragioni cronologiche. Basti qui accennare che la sua proposta è più orientata al sintattico (dove la nostra lo è al morfologico) e guarda più all’inglese (ed alla omologia con i tagset inglesi) che alla tradizione grammaticografica italiana (dove la nostra proposta è più sensibile alle esigenze della consensualità all’interno della tradizione italiana).

[16] Il tagset in uso a Stoccarda, lo “STTS” (Stuttgart/Tübinger Tagset), per il quale è anche disponibile un file di parametri per il TreeTagger, ne è una varietà (cfr. Schiller et alii 1995 e 1999) sviluppata da Anne Schiller (allora IMS/STR, ora RXRC/Grenoble), Christine Thielen (SfS/TÜB), Simone Teufel (allora IMS/STR, ora Cogsci/Edinburgh) e Christine Stöckert (IMS/STR), a partire dall'esperienza del corpus ELWIS (cfr. Hinrichs et alii 1995 e Feldweg - Kibiger - Thielen 1995).

[17] Cfr. i tagset presentati in Garside - Leech - Sampson 1987, appendice B.

[18] Per il tagset dell’ ICE (International Corpus of English) cfr. invece Greenbaum 1993.

[19] Cercando di avvicinarsi a «the ideal of providing distinct codings for all classes of words having distinct grammatical behaviour» (Garside - Leech - Sampson 1987, p. 167).

[20] Come, tra i corpora più recenti, lo IULA di spagnolo e catalano (cfr. Cabré et alii 1998).

[21] L’inglese (cfr. ad es. Leech 1997a, p. 25) rende possibile distinguere tra *tag* ‘categoria morfologica associata ad una determinata parola’ (ad esempio ‘preposizione’), *label* ‘il nome o la codifica con cui un *tag* è indicato’ (ad esempio “prep” o “IN”) e *adnotation* ‘l’operazione od il risultato dell’applicazione dei tag’ (ad esempio

Il CT, è vero, è stato etichettato semi manualmente e disambiguato con microregole; ma guardando più lontano, alla sua possibile estensione con tecniche stocastiche. Con il nostro tagset attualmente dimensionato a 67 tag (riducibili, alla bisogna, ad un minimo di 49) potremo, ad un costo descrittivo non poi troppo elevato, usare il CT come *training corpus* per annotare automaticamente con il TreeTagger (cfr. Schmid 1994) sviluppato dall'IMS altri testi italiani antichi, garantendo così un futuro scientifico ed una pubblica utilità alla nostra iniziativa. Il costo, si è detto, del contenimento del tagset non è molto elevato, perché è stato studiato in modo da essere ridotto al minimo. Lo strumento principale per ottenere ciò, come risaputo[22], è quello di alleggerire le informazioni già altrimenti codificate: distinzioni morfologiche "perse" a livello di tagset si possono recuperare scaricandole a livello lessicale[23] (ad es. nei pronomi).

Un altro esempio di limitazione computazionale è quello delle forme discontinue: dal momento che l'annotazione è attribuita ad ogni singola parola[24], non sono possibili tag compatti per i passivi ed i tempi composti. Tali categorie andranno gestite con regole di ricomposizione successive all'annotazione (*post-tagging rules*) ed elaborate a partire da essa. Il vantaggio indotto da questa "complicazione" è che presto avremo a disposizione delle regole ricavate da corpus da confrontare con quelle puramente "linguistiche" elaborate dai partecipanti ad *ItalAnt*.

2. LA STRUTTURA DI UN TAGSET: CARATTERISTICHE GENERALI**.** Se nei §§ 1.1-3 abbiamo esaminato quali siano i requisiti che un tagset deve soddisfare, vediamo ora a quali specifiche strutturali generali deve conformarsi, introducendo anche qualche indispensabile definizione.

2.1 *LABELS* E NOTAZIONI. Il sistema di "etichette" (*labels*) in cui si esprime un tagset è questione puramente convenzionale. L'importante è che tale sistema sia rigoroso e coerente in modo da consentire il mapping tra sistemi diversi con semplici procedure di conversione, vuoi per poter esportare informazioni in altre elaborazioni computazionali, vuoi per potere meglio eseguire particolari operazioni anche all'interno dello stesso progetto.[25]

Il sistema base di etichette che noi usiamo, e che trovate qui nelle tavole del tagset, è essenzialmente quello EAGLES, a base inglese (i puristi, al solito, inorridiranno), ma che ha l'indubbio vantaggio di essere immediatamente confrontabile con le altre descrizioni di tagset EAGLES, quali ELM-IT ed ELM-DE, alla maniera del documento Monachini - Calzolari 1996, alla cui copertura linguistica si può idealmente aggiungere. È questa quella che chiamiamo "notazione estesa" ("ExN" *Extended Notation*).

---

con_prep l'_art ombrello_n), laddove l'italiano dispone solo *di annotazione* ed *etichetta*. Io nel prosieguo cercherò di usare *etichetta* nel solo significato di 'label', ricorrendo a *tag* (in tondo: prestito non adattato) al posto di *annotazione* solo quando l'uso di *annotazione* nel senso di 'tag' riuscisse incongruo all'uso italiano o controindicato nel singolo contesto.

[22] Già Marcus - Santorini - Marcinkiewicz 1994, p. 274, infatti, scrivevano: «A key strategy in reducing the tagset was to eliminate redundancy by taking into account both lexical and syntactic information. Thus, whereas many POS tags in the Brown Corpus tagset are unique to a particular lexical item, the Penn Treebank tagset strives to eliminate such instances of lexical redundancy».

[23] Un esempio in cui questa strategia suona molto "naturale" dal punto di vista della tradizione linguistica italiana è la rinuncia ad introdurre uno specifico tag per il numero del possessore (oltre che per quello del posseduto, *nostro* vs *nostri*) nei possessivi, recuperandolo invece lessicalmente con i lemmi distinti *mio* e *nostro*.

[24] Tralasciando qui il problema, analogo, delle *multiword entries* (in italiano variamente chiamate "locuzioni", "unità multilessicali" o "polirematiche", su cui torneremo in séguito), già affrontato in altra sede (cfr. Barbera - Marello 2000). A proposito del quale basti qui dire che una possibile soluzione a livello di tagging è stata sperimentata nella più recente versione del CT (già online nel 2006), ma non era ancora stata sondata all'epoca dell'originaria comunicazione al convegno SILFI (2000).

[25] Importante è inoltre, come abbiamo scoperto in séguito con la pratica, è anche l'ottimizzazione delle *labels* ai fini della query, secondo le direttive che abbiamo impostato in Barbera ¶ 23, *infra*, ma di cui non avevamo ancora perfetta consapevolezza ai tempi in cui impostavamo il CT-Tagset, fissandolo poi nella attuale versione 1.3.

Oltre a questo sistema abbiamo anche un sistema numerico, che chiamiamo "notazione condensata" ("CdN" *Condensed Notation*) in cui tutte le ultime "foglie" di una gerarchia sono rappresentate da un unico codice "collassato" di tag[26]. Il concetto sarà più chiaro dopo che avremo introdotto la nozione di "gerarchia tipata", e per ora un esempio sarà più efficace di molte parole. Per la POS "nome" i codici "20" e "21" rappresentano rispettivamente `n.com` ("nome e comune") e `n.prop` ("nome e proprio"):

n		POS
com	prop	type
20	21	

Tav. 1: Gerarchia della POS *nome*: notazione estesa e condensata

Un terzo sistema di etichette, che chiamiamo "notazione breve" ("ShN" *Short Notation*), è quello che di fatto utilizzato come formalismo di interrogazione nel CWB[27], in cui per comodità di uso (le ricerche vengono infatti attuate con comandi da stringa, cioè intieramente scritti) la notazione estesa è stata ulteriormente abbreviata.

Riprendendo l'es. di cui alla Tav. 1, il *mapping* fra i 3 sistemi risulta il seguente:

ExN	CdN	ShN
n.com	20	n.c
n.prop	21	n.p

Tav. 2: Mapping tra i 3 tipi di notazione per la POS *nome*.

2.2 ANCORAMENTO MORFOLOGICO. Il tipo di annotazione che qui ci concerne nelle specifiche EAGLES è definito genericamente come "morfosintattico" proprio perché pur essendo di base morfologica, consente anche l'espressione subordinata di parametri sintattici[28] o comunque di altro livello di analisi.

Nella nostra annotazione l'ancoramento morfologico è stato reso più stretto, costituendo il default prevalente in caso di possibili alternative. Dato che in prospettiva computazionale, da un lato, la gestione un livello per volta è più semplice e, dall'altro lato, il particolare tipo di corpus che dobbiamo gestire è computazionalmente piuttosto complesso, ci è parso bene avanzare richieste di annotazione il più semplici, chiare ed omogenee possibili onde salvaguardare il massimo rendimento della procedura[29]. Al POS-tagging, quindi, perterrano le categorie prevalentemente morfologiche, ed a fasce successive di annotazione (in futuro sperabilmente imple-

[26] Questa notazione è quella che abbiamo usato internamente per annotare il corpus, perché (anche se può parere strano) è quella risultata più pratica (cioè più veloce e meno soggetta ad errore) nell'annotazione manuale.

[27] Un elenco completo del tagset in *Short Notation*, scritto come guida per l'interrogazione online del CT, è Barbera 2000/2006. Si tenga presente che è prevista una piccola revisione del sistema, in base alle esperienze di ricerca fatte in questi anni, che prevede piccole modifiche fatte per evitare coincidenze formali di etichette nelle ricerche con *wildcharacters*: *ind* nei pronomi, ad esempio, sarà sostituito con *idf* (per evitare la omografia con l'indicativo verbale), ed in generale si tenderà a sostituire le potenzialmente "pericolose" etichette monolittere con bilittere (ad es. *vb* per *v*, ecc.). Per i criteri di ciò, e per maggiori dettagli, cfr. oltre Barbera ¶ 23.

[28] Anzi, a livello di annotazione di lessico anziché di corpora, è possibile e consigliabile anche la specificazione di qualche caratteristica semantica: cfr. Monachini - Calzolari 1999, pp. 168-171.

[29] Volere troppo, a nostro giudizio, ci avrebbe portato a poco stringere.

mentabili) le altre categorie: quelle più propriamente sintattiche ad un chunking[30] e quelle semantiche e testuali alle rispettive annotazioni (e un elementare markup di tipo testuale è già stato implementato).

Dal nostro tagset è pertanto tendenzialmente esclusa ogni categoria solo semantica – come ad esempio l'aspetto (*label* `aspect`), peraltro già evitato in ELM-IT – o solo sintattica – come ad esempio la distinzione tra uso attributivo e non attributivo dell'aggettivo (*labels* `attr / nattr` in `adj`) e quella, più complessa, tra aggettivo pronominale e pronome. La distinzione tra congiunzioni subordinanti e coordinanti (*labels* `subord / coord` in `conj`) è probabilmente la più rilevante eccezione a questa strategia; in questa area ed aree limitrofe avevamo, peraltro, già dovuto rinunciare, sia pure a malincuore, alle "congiunzioni testuali" (`text`) ed agli "avverbi connettivi".

D'altra parte, le annotazioni di carattere testuale, cui pure molto teniamo, non possono, infatti, trovare adeguato spazio in questo strato di annotazione, ed andranno od introdotte caso per caso in fase di *post-tagging*, o, più opportunamente, pensate globalmente in un secondo tempo come una batteria separata che si appoggi alla precedente.

2.3 POST-TAGGING. Tutta una serie di operazioni che rimangono giocoforza fuori dall'annotazione, sono rimandate ad un momento successivo che si suole indicare come "*editing post-tagging*" o, più brevemente, "post-tagging" *tout court*.

Al di là di varie verifiche e ripuliture dei dati (verifiche di correttezza del formario e disambiguazione delle forme per le quali più tag sono possibili), in questa fase si possono recuperare alcune distinzioni grammaticali (di natura sintattica, testuale e semantica già parzialmente previste) escluse dal tagset[31].

3. LA STRUTTURA DI UN TAGSET: LE GERARCHIE TIPATE**.** Abbiamo più volte accennato alla natura essenzialmente gerarchica del nostro tagset, così come dei tagset EAGLES-conformi e dei tagset usati nella linguistica dei corpora in genere.

Per meglio spiegarci usiamo un caso concreto: per la POS (*Part Of Speech*) "nome" la procedura GATTO del Padua Corpus ereditata dall'OVI (cfr. Barbera - Marello 1999/2001: § 5) usa prevalentemente tre tag separati, etichettati `sm sf` e `np` (il cui valore è facilmente immaginabile), oltre a tutta una serie di tag meno frequenti (come `ng` per i nomi geografici), laddove il nostro disegno prevede un unico tag che si identifica con la POS "nome" ed etichettato `n`, che si suddivide in due *types* (ossia "tipi", donde la nozione di "tipato"), etichettati `com` e `prop`, che potrebbero poi ulteriormente ramificarsi in più *features* e *sub-features*. Ipotizzando di voler trovare tutte le sequenze di "nome_aggettivo" in un sistema ad etichette gerarchiche possiamo cercare semplicemente "`n_adj`", laddove in un sistema ad etichette compatte come quello di GATTO dovremmo usare una lunga catena di congiunzioni, "`sm&sf&np&ng&..._agg`".

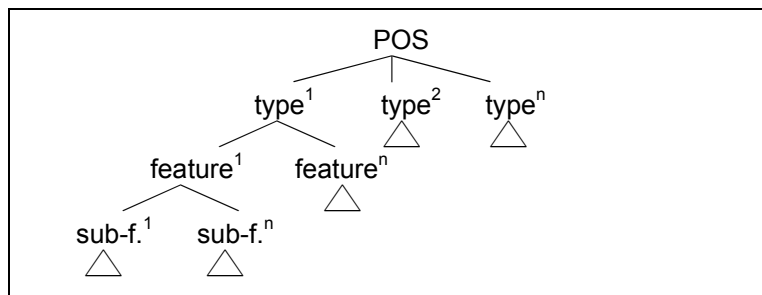
L'utilizzo, ossia, di etichette analitiche nella annotazione di un corpus ne permette una descrizione dettagliata e ricerche specifiche, ma l'analiticità risulta dispersiva ed impedisce ricerche generali se non viene sussunta in un sistema di generalizzazioni gerarchiche, fondata sull'ereditarietà.

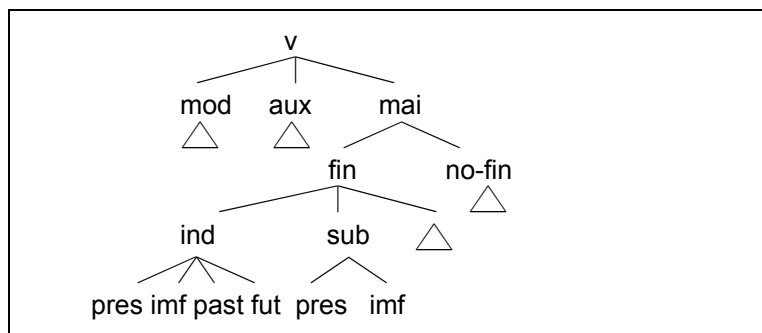
---

[30] Ad un vero parsing non abbiamo mai pensato, vuoi per scarso convincimento teorico (al più penserei ad uno *shallow parsing*), vuoi per difficoltà pratiche. Sono in effetti in corso sperimentazioni con l'ottimo chunker ricorsivo dell'IMS Stuttgart, lo YAC (cfr. Kermes - Evert 2002).

[31] Ma per le multiword cfr. qui sopra nota 24.

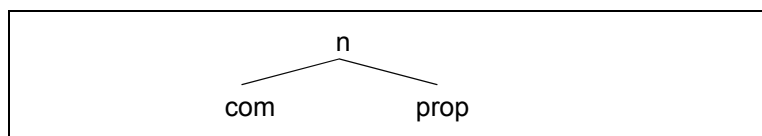
3.1 HDF E GERARCHIE TIPATE. Per facilitare la valutazione dell'esempio precedente abbiamo parlato di POS che si suddividono in *types* e quindi in *features* e *sub-features*. In realtà l'approccio definitorio di EAGLES procede piuttosto in senso contrario, *bottom-up*: si parla così di gruppi di *hierarchy-defining features* (HDF), di annotazioni, cioè, che si costruiscono in una gerarchia, e non viceversa.

In altri termini, tutte le POS sono la proiezione di un fascio di *features* gerarchiche (HDF); il loro *branching* più alto è detto `Type`[32] ed i *sub-branchings* via via più bassi sono le *features* POS-specifiche (*subfeatures*). Dal punto di vista puramente computazionale, comunque, la questione del verso (*bottom-up* o *top-down*) non è rilevante, in quanto le gerarchie tipate sono percorribili indifferentemente in entrambi le direzioni.

Tav. 3: Schema arborescente di una classe di HDF.

La tavola precedente riproduce lo schema arborescente[33] di una "classe di HDF" (che per brevità conveniamo di chiamare semplicemente "HDF"). Ed illustrerò ulteriormente l'argomento, data la sua importanza, con due esempi concreti, il "verbo" ed il "nome". Il primo offre un esempio di HDF altamente ramificanti (e per fortuna nel nostro tagset è il solo caso di tale complessità),

Tav. 4: Schema arborescente della classe HDF "verbo".

[32] Da non confondersi con il "type" relato con "token": l'uno (il *type* gerarchico) lo consideriamo termine talmente specifico da potersi mantenere in inglese (e pertanto con plurale in *-s* e stampato in corsivo), laddove l'altro lo trattiamo come prestito non adattato (plurale invariabile e stampato in tondo).

[33] Oltre al diagramma ad albero, sono naturalmente allo stesso titolo possibili anche altre modalità di rappresentazioni (ad esempio a matrice, ad arco, od a blocchi). Si tratta, infatti, comunque di «oggetti astratti e distinti dalla loro rappresentazione tipografica» (Allegranza - Mazzini 2000, p. 146).

ed il secondo di scarsamente ramificanti (e nel nostro tagset la gran parte dei casi sono proprio così):

Tav. 5: Schema arborescente della classe HDF "nome".

3.2 MSF E *CROSS-BRANCHING*. Abbiamo visto come trattare le *features* che si proiettano su una gerarchia risalendo alla POS lungo un unico percorso. Ma non tutte hanno queste caratteristiche. Il genere, ad es., non risale ad un'unica sorgente, ma si proietta bensì su più POS o tipi distinti (nome, aggettivo, pronome, participio). In altre parole, gli alberi che descrive si incrociano con molteplici *cross branching*, vanificando la inequivocità dell'ereditarietà gerarchica.

Bisogna pertanto distinguere alcune MSF (*morphosyntactic features*) dalle HDF (*hierarchy-defining features*). Nell'architettura EAGLES (e pertanto nella nostra) solo le seconde si costruiscono in gerarchia tipata, mentre le prime si applicano liberamente sui tag tipati.

Tutte le MSF ammettono una sola classe di valori (`values`) alternativi: in altre parole non presentano *sub-branching* di sorta. Ad esempio per il numero se ne hanno solo tre: `numb {sg, pl, n}`, e così via.

Non c'è sempre universale consenso che una classe di tratti alternativi debba venire considerata una *sub-branched* HDF od una MSF: così `VfM` (*verbal forms \ moods*) è gerarchica per ELM-DE ma non-gerarchica per ELM-IT. A prescindere da considerazioni di comodità informatica, la coerenza vorrebbe che, comunque, fosse gerarchizzata ogni classe di tratti alternativi che sia POS-specifica (ma `tns` "tempo", che pure compare solo nel verbo è trattato come MSF tanto da ELM-IT, ELM-DE e dai sistemi descritti in MORPHSYN) e che fosse mantenuta come MSF ogni classe di tratti alternativi che si applichi a più POS (ma `degr` "grado", comune ad `adj` ed `adv` è da tutti trattato come HDF).

4. DICHIARAZIONE PROGRAMMATICA. Il percorso per arrivare al CT-tagset, a partire da queste considerazioni generali, è lungo e frastagliato, e sarà forse utile farlo precedere da quella dichiarazione programmatica che avevo steso ancora nel 1999[34], all'inizio di questa avventura: ché forse, come aveva guidato noi allora, potrebbe oggi fare da guida anche al lettore. Si tratta di una sorta di decalogo, più prescrittivo che argomentativo: tutte le nozioni cui fa cenno sono comunque variamente discusse altrove in questo articolo.

I principali fattori che abbiamo deciso di tenere presenti nella costruzione del tagset sono:

(j) Massima compatibilità con il tagset tedesco dell'IMS di Stoccarda e le (in larga parte coerenti) raccomandazioni di EAGLES. Esportabilità verso una nuova suite di tagset miltilingui costruiti secondo la medesima struttura.

(ij) Tentativo di contenimento del numero dei tag nell'eventualità dell'introduzione di procedimenti stocastici. La quantità dei tag HDF (cfr. *infra* per la definizione della nozione) "compositi" totali deve essere inferiore a 70 (36 sono previste nel solo verbo) per un tagger stocastico. Scartata è l'ipotesi di una ulteriore riduzione preliminare[35]: a sfrondare un tagset, se necessario, c'è sempre tempo.

---

[34] E che riproduco qui sostanzialmente invariata dal documento interno che la conteneva.

[35] L'eliminazione, ad es., di tag verbali come `tns` e `mod`, ridurrebbe drasticamente il numero delle combinazioni complessive, ma ridurrebbe di molto l'efficacia di molte regole.

(iij) Prevalente ancoramento, per pulizia del sistema, dei tag “morfosintattici” al livello morfologicamente esplicito. È pertanto tendenzialmente esclusa ogni categoria semantica (e.g. `aspect`, ecc.), come già in ELM-IT, o puramente sintattica (e.g. `attr / nattr` in `adj`, distinzione tra aggettivo pronominale e pronome, ecc.).

(iiij) Rinvio al *post-tagging* di un certo numero di tag di natura sintattica, testuale e semantica già parzialmente previsti, così come delle forme verbali composte.

Le nozioni generali presupposte dalla dichiarazione di questo tagset sono quindi le seguenti:

(1) Si distingue tra **HDF** (*hierarchy-defining features*) e **MSF** (*morphosyntactic features*). Nella struttura tanto di EAGLES quanto di IMS (e pertanto nella nostra) le prime si costruiscono in gerarchia tipata, le seconde no e si applicano liberamente sui tag tipati.

(2) Tutte le **POS** (*part of speech*) hanno *features* gerarchiche (HDF) ed il loro primo *sub-branching* è detto *`Type`*; i successivi *sub-branching* sono POS-specifici (*subfeatures*).

(3) Le MSF ammettono una sola classe di *`Values`* alternativi (`numb {sg, pl, n}`, ecc.).

(4) Anche se, come s’è detto, non v’è sempre accordo sulla questione, nel sistema proposto la ripartizione tra MSF e HDF sarà rigorosa (sono HDF `solo` *features* che ereditano un’unica POS) in modo da evitare *sub-branching* incrociati.

5. IL CT-TAGSET**.** Il sistema complessivo così disegnato, tenuto conto delle raccomandazioni EAGLES e di tutte le considerazioni esposte nei paragrafi precedenti, comprende complessivamente cinque MSF e dodici HDF, secondo riassunto nella tavola seguente:

HDF	(1) *noun*, (2) *verb*, (3) *adjective*, (4) *pro-det*, (5) *adverb*, (6) *conjunction*, (7) *adposition*, (8) *article*, (9) *numeral*, (10) *interjection*, (11) *punctuation*, (12) *residual*
MSF	(1) *person*, (2) *gender*, (3) *number*, (4) *degree*, (5) *multiword*

Tav. 6: Le *features* gerarchiche e morfologiche del CT Tagset.

Per comodità di etichettatura ad ogni *value* di MSF ed ad ogni foglia terminale di HDF sarà assegnato un codice numerico univoco. Si avranno pertanto (come illustrato in § 2.1), già in partenza due sistemi notazionali distinti e complementari: una notazione estesa (**ExN**) ed una notazione condensata (**CdN**).

Nella assegnazione delle etichette sono date prima le HDF e poi, in ordine fisso, le MSF (cfr. il “bastone” descritto al § 6.2), ma per ragioni di perspicuità nella descrizione dettagliata qui sotto fornita (al cui ordine è anche parzialmente ancorato quello dei codici numerici) presenteremo prima le MSF e poi le HDF.

5.1 LE *MORPHOSYNTACTIC FEATURES* (MSF). Vediamo ora più nei dettagli a definizione delle cinque MSF, con i loro valori e codici numerici.

5.1.1 MSF PERSON. Questo lo schema generale per la prima MSF:

MSF	**1**	***person***		
		feature	*value*	*code*
		`pers`	`1`	`1`
			`2`	`2`
			`3`	`3`

Tav. 7: La *feature* morfosintattica (MSF) `person`.

Non vi sono macro specifici per le disgiunzioni: una forma di congiuntivo presente singolare sarà pertanto `1;2;3`.

5.1.2 MSF GENDER. Questo lo schema generale per la seconda MSF:

MSF	2	***gender***		
		feature	*value*	*code*
		gend	masc	4
			fem	5
			c	4;5

Tav. 8: La *feature* morfosintattica (MSF) `gender`.

Si noti che `c` = *common* era stato introdotto come semplice macro della disgiunzione `masc;fem`, e non come tag autonomo[36].

5.1.3 MSF NUMBER. Questo lo schema generale per la terza MSF:

MSF	3	***number***		
		feature	*value*	*code*
		numb	sg	6
			pl	7
			n	6;7

Tav. 9: La *feature* morfosintattica (MSF) `mumber`.

Qui `n` = *invariant* è stato introdotto come macro della disgiunzione `sg;pl`.

5.1.4 MSF DEGREE. Questo lo schema generale per la quarta MSF:

MSF	4	***degree***		
		feature	*value*	*code*
		degr	pos	8
			comp	9
			sup	10

Tav. 10: La *feature* morfosintattica (MSF) `degree`.

La maggior parte dei tagset EAGLES gestiscono il grado come HDF, ma la *feature* è comune ad `adj` ed `adv`, sicché qui si è preferito evitare ogni possibile *cross-branching*. I `comp \ sup` analitici, poi, vanno trattati con *multiword expression tagging*[37]: questo comporta che saranno etichettati come `10` solo i superlativi assoluti, mentre i relativi avranno il tag `9` (comparativo).

[36] E di fatto, poi, nel prosieguo della annotazione non è stato mai utilizzato.
[37] Cfr. ELM-IT che rimanda ad un introvato Leech & Wilson, *Invitation Draft*.

5.1.5 MSF MULTIWORD. Il trattamento, più volte accennato, che abbiamo sperimentato per le locuzioni (od unità polirematiche o *multiword*, all'occorrenza abbreviate con la sigla "MW"), si basa su una marca (introdotta fin dalle prime fasi della ricerca) di MSF. Questo lo schema generale per la così costituita quinta MSF:

MSF	**5**	***multiword***		
		feature	*value*	*code*
		loc	const	11
			two	12
			three	13
			n ≤ nine	1*n*

Tav. 11: La *feature* morfosintattica (MSF) *multiword*.

In linea di massima, infatti, le MW sono trattabili come MSF perché, comunque, si distribuiscono su più POS. Sono previste dal sistema locuzioni costituite da due a nove costituenti. Il *value* const (*constituent*, 11) è attribuito alle singole parti costituenti la polirematica[38].

5.2 POS E *HIERARCHY DEFINING FEATURES* (HDF). Nel prosieguo presenteremo tutte le gerarchie tipate di tutte le POS del tagset del *Corpus Taurinense*, in duplice versione tabulare (schema generale e schema gerarchico), e con una discussione minima dei criteri che ne hanno ispirato la costruzione.

5.2.1 LA POS NOME ("NOUN" = "N": 2 TAG). La POS costruita per i nomi è molto semplice. Questo lo schema generale:

HDF 1 ***noun*** (*2 comp. HDF tags*)		+ MSF gend, numb, loc
POS	*types*	
n	com	
	prop	

Tav. 12: La HDF *noun*: schema generale

Volendo, ulteriori distinzioni (variamente tradizionali e raccolte in Monachini - Calzolari 1996) potrebbero essere introdotte in *post-tagging*. Questa la tavola riassuntiva:

n		POS
com	prop	type
20	21	

Tav. 13: La HDF *noun*: schema gerarchico

[38] Di solito il *value* const viene attribuito in associazione ad un lemma che può avere qualsiasi HDF (nel caso che almeno una sua forma sia attestata anche al di fuori della sola polirematica, o che faccia comunque parte di una POS chiaramente individuata, ad es. un nome proprio) od una POS zero (nel caso di costituenti che ricorrano esclusivamente nella polirematica in esame e che non possano così essere automaticamente attribuiti ad una POS specifica).

5.2.2 LA POS VERBO ("VERB" = "V": 36 TAG). La POS disegnata per i verbi è, come già accennato, di gran lunga la più complessa e gerarchicamente articolata del CT tagset

Lo schema generale è infatti il seguente:

HDF 2 ***verb*** (*36 comp. HDF tags*)					
POS	*types*	*fin*	*VfMs*	*tns*	+ MSF
v	mai	fin	ind	pres	pers, gend, numb, loc
	aux	no-fin	sub	ipf	
	mod		cond	past	
			impr	fut	
			inf		
	part				
			ger		

Tav. 14: La HDF `verb`: schema generale

Questa la tavola gerarchica, rimpicciolita e spezzata in tre per ragioni di spazio:

<table>
<tr><td colspan="16">v</td><td>POS</td></tr>
<tr><td colspan="14">mai</td><td>aux</td><td>mod</td><td>type</td></tr>
<tr><td colspan="8">fin</td><td colspan="6">no-fin</td><td></td><td></td><td>fin</td></tr>
<tr><td colspan="2">ind</td><td colspan="2">sub</td><td colspan="2">cond</td><td colspan="2">impr</td><td>inf</td><td>121</td><td colspan="2">part</td><td colspan="2">ger</td><td>...</td><td>...</td><td>Vfm</td></tr>
<tr><td>pres</td><td>111</td><td>pres</td><td>115</td><td>pres</td><td>117</td><td>pres</td><td>118</td><td></td><td></td><td>pres</td><td>122</td><td>pres</td><td>124</td><td></td><td></td><td rowspan="4">tns</td></tr>
<tr><td>ipf</td><td>112</td><td>ipf</td><td>116</td><td></td><td></td><td></td><td></td><td></td><td></td><td>past</td><td>123</td><td></td><td></td><td></td><td></td></tr>
<tr><td>past</td><td>113</td><td colspan="14"></td></tr>
<tr><td>fut</td><td>114</td><td colspan="14"></td></tr>
</table>

<table>
<tr><td colspan="15">v</td><td></td><td>POS</td></tr>
<tr><td>mai</td><td colspan="14">aux</td><td>mod</td><td>type</td></tr>
<tr><td></td><td colspan="8">fin</td><td colspan="6">no-fin</td><td></td><td>fin</td></tr>
<tr><td>...</td><td colspan="2">ind</td><td colspan="2">sub</td><td colspan="2">cond</td><td colspan="2">impr</td><td>inf</td><td>221</td><td colspan="2">part</td><td colspan="2">ger</td><td>...</td><td>Vfm</td></tr>
<tr><td></td><td>pres</td><td>211</td><td>pres</td><td>215</td><td>pres</td><td>217</td><td>pres</td><td>218</td><td></td><td></td><td>pres</td><td>222</td><td>pres</td><td>224</td><td></td><td rowspan="4">tns</td></tr>
<tr><td></td><td>ipf</td><td>212</td><td>ipf</td><td>216</td><td></td><td></td><td></td><td></td><td></td><td></td><td>past</td><td>223</td><td></td><td></td><td></td></tr>
<tr><td></td><td>past</td><td>213</td><td colspan="13"></td></tr>
<tr><td></td><td>fut</td><td>214</td><td colspan="13"></td></tr>
</table>

<table>
<tr><td colspan="16">v</td><td>POS</td></tr>
<tr><td>mai</td><td>aux</td><td colspan="14">mod</td><td>type</td></tr>
<tr><td>...</td><td>...</td><td colspan="8">fin</td><td colspan="6">no-fin</td><td>fin</td></tr>
<tr><td></td><td></td><td colspan="2">ind</td><td colspan="2">sub</td><td colspan="2">cond</td><td colspan="2">impr</td><td>inf</td><td>321</td><td colspan="2">part</td><td colspan="2">ger</td><td>Vfm</td></tr>
<tr><td></td><td></td><td>pres</td><td>311</td><td>pres</td><td>315</td><td>pres</td><td>317</td><td>pres</td><td>318</td><td></td><td></td><td>pres</td><td>322</td><td>pres</td><td>324</td><td rowspan="4">tns</td></tr>
<tr><td></td><td></td><td>ipf</td><td>312</td><td>ipf</td><td>316</td><td></td><td></td><td></td><td></td><td></td><td></td><td>past</td><td>323</td><td></td><td></td></tr>
<tr><td></td><td></td><td>past</td><td>313</td><td colspan="12"></td></tr>
<tr><td></td><td></td><td>fut</td><td>314</td><td colspan="12"></td></tr>
</table>

Tav. 15: La HDF `verb`: schema gerarchico

Le *features* aspect {perf, imperf}, voice {act, ps}, refl {...}, MaiVF {trans, intrans, imp} previste da MORPHSYN non sono supportate né da ELM-DE né da ELM-IT, né tantomeno lo sono da noi[39]. V è l'unica POS ad avere un *sub-branching* molto pesante: anche per questa ragione si è preferito mantenere a livello lessicale e non di tag la marca di "pronominalità", in ciò, peraltro, secondando l'originaria impostazione dell'OVI.

Per maggiore perspicuità, data la consistenza numerica del sistema, si è scelto di attribuire ai tag verbali codici di tre cifre: la prima indica il type{1=mai;2=aux;3=mod}, la seconda la *finiteness* {1=fin;2=no-fin} e la terza {1-8} le varie combinazioni di VfM e tns.

5.2.3 LA POS AGGETTIVO ("ADJECTIVE" = "ADJ": 1 TAG). Semplicissima invece la POS costruita per gli aggettivi:

HDF 3 ***adjective*** (*1 comp. HDF tag*)		
POS	*types*	
adj	qual	+ MSF gend, numb, degr

Tav. 16: La HDF *adjective*: schema generale

In questa ipotesi riduzionistica sono pertanto adj solo i qualificativi. L'ulteriore *feature* use, per distinguere l'uso predicativo dall'attributivo è da rimandare al *post-tagging*; il problema degli aggettivi pronominali[40] è invece considerato nella POS successiva.

adj		POS
qual	26	type

Tav. 17: La HDF *adjective*: schema gerarchico

5.2.4 LA POS PRONOME-DETERMINANTE ("PRO-DET" = "PD": 11 TAG). La costruzione di una sola POS per pronomi e determinanti, due gruppi di forme già tradizionalmente affatto eterogenei già al loro interno, è di quelle che hanno costato molta riflessione. Lo schema che presentiamo qui sotto è il risultato delle riflessioni svolte in Barbera 2000/2003:

HDF 4 ***pro-det*** (*11 comp. HDF tags*)				
POS	*types*	*infl*	*cases*	
pd	dem indf poss int rel pers excl	weak strg	nom obl	+ MSF pers, gend, numb

Tav. 18: La HDF *pro-det*: schema generale

[39] Della discordanza nel trattamento di VfM e tns come HDF anziché MSF si è già detto poco sopra; si ricorda anche che i tempi composti ed i passivi sono da ricavare con apposite regole di *post-tagging*.

[40] Il *type* det, infatti, è da introdurre solo se si vuole trattare così i "pronomi" aggettivali, o "determinanti".

Si tratta, in pratica, di una classe arbitraria a definizione morfologica, per la cui giustificazione rimandiamo all'articolo citato[41].

Accantonando, dunque, i rapporti inter-POS tra "aggettivi", "pronomi" ed "avverbi", restano da documentare alcune scelte puntuali concernenti alcune *(sub)features*.

La *feature* `weak` di `dem` è stata introdotta per coprire parte delle cosiddette "particelle"[42].

La *feature* `weak` nei `poss` è stata introdotta per notare il tipo *soreta*, la cui estensione per quanto modesta è comunque superiore all'odierna (cfr. il pisano *suorse* 'le sue sorelle' riportato in Rohlfs 1966-69, § 430 pp. 124-5, da Castellani 1965, p. 134). Ancora per quanto riguarda i `poss`, un altro problema si ha con l'intreccio tra numero del possessore e del posseduto: usando una sola *subfeature* nella dichiarazione HDF e le sole MSF qui introdotte, infatti, *tuo* e *vostro* si trovano ad avere una sola etichetta (`33,2,4,6,0,0`); la distinzione è comunque recuperabile dalla associazione lemmatica diversa, giusta la tecnica illustrata nel § 1.3 e nota 22.

Quanto, invece, ai `pers`, le maggiori difficoltà si incontrano alla *subfeature* `case`, dove il *value* `obl` raccomandato da ELM-IT è nettamente ipodifferenziato. In realtà (come parzialmente riconosciuto anche in ELM-IT) andrebbero distinti `nom;acc;prep` per l'*inflection* `strg` ed `acc;dat;eth` per l'*inflection* `weak` (cfr. es. come *dimmeglielo* in cui nei clitici si hanno in successione *ethic-dative-accusative*). In ottica riduzionista si è tenuto il *value* `obl` ipodifferenziato

Questa pertanto la tavola gerarchica riassuntiva della POS `pro-det`:

<table>
<tr><td colspan="12">P-D</td><td>POS</td></tr>
<tr><td colspan="2" rowspan="2">dem</td><td>indf</td><td colspan="2" rowspan="2">poss</td><td>int</td><td>rel</td><td colspan="4" rowspan="2">pers</td><td>excl</td><td rowspan="2">type</td></tr>
<tr><td>32</td><td>35</td><td>36</td><td>40</td></tr>
<tr><td>strg</td><td>weak</td><td></td><td>strg</td><td>weak</td><td></td><td></td><td colspan="2" rowspan="2">strg</td><td colspan="2" rowspan="2">weak</td><td></td><td rowspan="2">infl</td></tr>
<tr><td>30</td><td>31</td><td></td><td>33</td><td>34</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>nom</td><td>obl</td><td>nom</td><td>obl</td><td></td><td rowspan="2">case</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>37</td><td>38</td><td>41</td><td>39</td><td></td></tr>
</table>

Tav. 19: La HDF *`pro-det`*: schema gerarchico

5.2.5 LA POS AVVERBIO ("ADVERB" = "ADV": 2 TAG). Decisamente riduzionista è la struttura adottata per la POS avverbio, altra categoria, come i pronomi, linguisticamente del tutto eterogenea. Per i rapporti con il problema generale dei `pd` cfr. *supra* (e nel senso specificato lì va letta la mancata introduzione della *feature* `wh`). In ottica morfologico-riduzionista, poi, è inevitabile la rinuncia alle categorie come `fras` (cfr. *supra*). È stato però introdotto per i clitici *ci*, *ne*, *vi* con valore neutro-locativo il type `particle`[43]. E dunque:

---

[41] Di fatto, ciò si traduce nel rimandare la distinzione funzionale tra `pro` / `adj` / `adv` al *post-tagging* o ad altre strategie.

[42] In particolare si sono sempre distinti tre principali tipi di *ne*: (1) "pronominale", a valore dimostrativo (e.g. *dàmme<u>ne</u>*, ecc.), etichettato "`p-d.det.weak`"; (2) "avverbiale", a valore neutro o locativo (e.g. *vàtte<u>ne</u>*, *se <u>ne</u> va*, ecc.), etichettato "`adv.particle`"; (3) "personale", equivalente ad 'a noi' (e.g. *<u>ne</u> dice*, ecc.), etichettato "`pers.weak.obl`". Analogamente, a "`p-d.det.weak`" devono inoltre essere ricondotti anche i *ci*, *vi* a valore dimostrativo (e.g. *non <u>ci</u> credo*), mentre – come ovvio – quelli a valore personale (e.g. *non <u>ci</u> conviene*) andranno etichettati come "`pers.weak.obl`", e quelli a valore "avverbiale" neutro o locativo (e.g. *non <u>ci</u> entra*), riceveranno invece il tag "`adv.particle`" (indipendentemente da quale potrà essere la scelta in sede di *post-tagging* per i verbi come *esserci*).

[43] Per cui cfr. nel § 5.2.4 sui `pro-det` e nota 42.

HDF 5 ***adverb*** (*2 comp. HDF tags*)		
POS	*types*	+ MSF `degr, loc`
`adv`	`general` `particle`	

Tav. 20: La HDF `adverb`: schema generale

Questa infine la tavola riassuntiva:

`adv`			`POS`
`general`	`particle`	`(...)`	`type`
`45`	`46`		

Tav. 21: La HDF `adverb`: schema gerarchico

5.2.6 LA POS CONGIUNZIONE (“CONJUNCTION” = “CONJ”: 2 TAG). Altrettanto riduzionista e spartana è pure la POS congiunzione:

HDF 6 ***conjunction*** (*2 comp. HDF tags*)		
POS	*types*	+ MSF `loc`
`conj`	`coord` `subord`	

Tav. 22: La HDF `conjunction`: schema generale

Da un lato, la granularità con i soli `coord;subord` è certo scarsa, dall’altro però già la consistenza stessa della POS è sintatticamente “sporca” (il discrimine verso le adposizioni riducendosi di fatto acché le prime sono introduttori di frasi, le seconde di sintagmi nominali): la coerenza con le direttive EAGLES (dove sono sempre distinte) ed il rispetto della tradizione grammaticale italiana, però, hanno reclamato il loro conto. Questa la tavola riassuntiva:

`conj`			`POS`
`coord`	`subord`	`(...)`	`type`
`50`	`51`		

Tav. 23: La HDF `conjunction`: schema gerarchico

5.2.7 LA POS ADPOSIZIONE (“ADPOSITION” = “ADP”: 2 TAG). Non problematica (salvo quanto osservato sopra in § 5.2.6) la POS adposizione:

HDF 7 ***adposition*** (*2 comp. HDF tags*)		
POS	*types*	+ MSF `loc`
`adp`	`prepos` `postpos`	

Tav. 24: La HDF `adposition`: schema generale

Le preposizioni articolate sono state gestite in fase di tokenizzazione, e sono quindi etichettate con tag separati[44] e notazione di grafoclisia (es. *a ÷lle*, con due token etichettati separatamente) per distinguerle dalle separate (tanto più che la questione in italiano antico a volte è più che altro editoriale).

Il tag `postpos` (assente in ELM-IT) è qui introdotto per i vari *meco*, *teco*, ecc. (tokenizzati *me ÷co*, *te ÷co*). Questa la tavola riassuntiva:

adp		POS
prepos	postpos	type
56	57	

Tav. 25: La HDF `adposition`: schema gerarchico

5.2.8 LA POS ARTICOLO ("ARTICLE" = "ART": 2 TAG). Eluse le istanze sintattiche (che muoverebbero in direzione determinante) a favore della omostrutturalità con gli altri tagset EAGLES e della volontà di non rompere con la tradizione grammaticale italiana, la struttura della POS articolo appare abbastanza scontata:

HDF 8 ***article*** (*2 comp. HDF tags*)		
POS	*types*	+ MSF gend, numb
art	def indef	

Tav. 26: La HDF `article`: schema generale

art		POS
def	indef	type
60	61	

Tav. 27: La HDF `article`: schema gerarchico

5.2.9 LA POS NUMERALE ("NUMERAL" = "NUM": 2 TAG). Tradizionale[45] la struttura anche di questa POS:

HDF 9 ***numeral*** (*2 comp. HDF tags*)		
POS	*types*	+ MSF gend, numb
num	card ord	

Tav. 28: La HDF `numeral`: schema generale

Ed eccone la tavola riassuntiva:

[44] Quindi niente `form=fuse`, come pur sarebbe possibile in ottica EAGLES. Cfr oltre n. 55.

[45] Che linguisticamente i numerali siano dei quantificatori, come anche gran parte dei tradizionali indefiniti è indubbio; il danno linguistico, almeno nell'ottica di strategie di query sul corpus etichettato, non è tuttavia forte.

num		POS
card	ord	type
64	65	

Tav. 29: La HDF `numeral`: schema gerarchico

5.2.10 LA POS INTERIEZIONE ("INTERJECTION" = "INTJ": 1 TAG). Nulla da rimarcare se non la ovvia rinuncia ad una classificazione semantica:

HDF 10 ***interjection*** (*1 comp. HDF tags*)		
POS	*types*	+ MSF
intj	-	loc

Tav. 30 La HDF `interjection`: schema generale

intj	POS
general	type
68	

Tav.31: La HDF `interjection`: schema gerarchico

5.2.11 LA POS PUNTEGGIATURA ("PUNCTUATION" = "PUNCT": 2 TAG). La punteggiatura, assente in ELM-IT, è stata messa dall'ELM-DE tra i `resid`; qui si è invece preferito assegnarle[46] una POS autonoma, la cui struttura è bipartita:

HDF 11 ***punctuation*** (*2 comp. HDF tags*)		
POS	*types*	[Ø MSF]
punct	fin non-fin	

Tav. 32 La HDF `punctuation`: schema generale

I singoli interpuntemi sono pertanto trattati come entrate lessicali, lemmatizzate con i loro nomi tipografici internazionali[47]. Le *feature values* `fin` e `non-fin`, inoltre, dato che nel corpus su cui operiamo non sono sempre distinguibili le partizioni testuali al di sopra del periodo (accapo, sezione, paragrafo), devono intendersi come aventi dominio d'applicazione il solo periodo (una virgola sarà, pertanto, sempre `no-fin` ed un punto `fin`).

punct		POS
fin	nonfin	type
70	71	

Tav.33: La HDF `punctuation`: schema gerarchico

[46] Come peraltro possibile negli schemi EAGLES.
[47] E cioè come <,> *comma*, <:> *colon*, <;> *semicolon*, <.> *stop*, <–> *emdash*, <…> *ellipsis*, <!> *exclam*, <?> *question*, <'> *quote*, <«> *guillemotleft*, <»> *guillemotright*, <(> *parenleft*, <)> *parenright*, <"> *quotedouble*, ecc.

5.2.12 LA POS "RESIDUI" ("RESIDUAL" = "RES": 4 TAG). I *types* raccolti in questa gerarchia ("*wastebasket-hierarchy*") sono inerentemente eterogenei:

HDF 12 ***residual*** (*3 comp. HDF tags*)		
POS	*types*	
resid	frgn abbr formula epenth	+ MSF gend, num, loc

Tav. 34 La HDF `residual`: schema generale

Il trattamento delle *foreign words* è tanto in ELM-IT quanto in ELM-DE; le *abbreviations* in ELM-DE sono solo `trunc` (che sono altra cosa: primo membro di composto)[48]. Anche i simboli grafici ({SC} '*signum crucis*', ecc.) e filologici (* '*vacuum*', × '*lacuna*' e ^ '*deperditum*') sono stati marcati `abbr`; *e converso*, si è stati abbastanza avari ad assegnare ad `abbr` forme attribuibili ad una esplicita classe morfosintattica e/o lemma pieno, restringendo il tag alle sole abbreviazioni fortemente convenzionalizzate (come *etc~*[49]), di valore incerto, od alle unità di misura (tipo *l~, den~,* ecc.) della cui forma piena la valenza linguistica (genere e numero) è spesse volte molto vaga.

Il *type* `formula`, introdotto sulla base della maggior parte dei tagset EAGLES per qualsiasi notazione numerica e non linguistica di espressioni numerali, si è poi rivelato poco efficace, almeno in questo particolare tipo di corpus[50]. Il *type* `epenth` raccoglie le particelle epentetiche (o paragogi) *-e* e *-no*, che si desiderava poter studiare in modo più puntuale[51].

Questa la tavola riassuntiva:

res				POS
frgn	abbr	formula	epenth	type
75	76	77	78	

Tav.35: La HDF `residual`: schema gerarchico

6. *FEATURE DECLARATIONS* (FD) E MAPPING INTERNOTAZIONALE. Forniamo in questo ultimo blocco di paragrafi una prospezione generale della "features declaration" prevista dal nostro sistema, unitamente ad un *mapping* tra le nostre tre diverse notazioni (§ 6.1), ad una tavola delle associazioni obbligatorie tra HDF e MSF (§ 6.3), ed ad una presentazione schematica del nostro formato complessivo di annotazione (il cosiddetto "bastone di annotazione": § 6.2).

[48] Una riflessione che ci è stato dato di fare, purtroppo, solo a corpus finito, quando i giochi erano ormai fatti, è che `abbr` (e forse anche `frgn`) sarebbe forse più utilmente stato introdotto come MSF: per una futura versione del *Corpus Taurinense* (e per futuri tagset ispirati alla sua struttura) è questo un punto su cui potrebbe valer la pena di tornare sui nostri passi.

[49] Il segno <~> è un sostituto convenzionale del punto abbreviativo introdotto in fase di tokenizzazione per evitare la collisione omografica tra punto interpuntivo ed abbreviativo.

[50] Anche se nella versione finale del CT risulta di fatto inutilizzato, si è mantenuto nello schema generale di annotazione, pensando a testi futuri che contengano intiere espressioni puramente numeriche (cioè completi chunks non linguistici).

[51] Anche questo tag non è risultato particolarmente indispensabile linguisticamente, e potrà eventualmente essere eliminabile in futuro.

6.1 LA DICHIARAZIONE DELLE HDF E DELLE MSF. Nelle due tavole seguenti è riportata la FD del CT tagset, documento indispensabile per ogni tagset tipato.

Per le *MSFeatures* è riportata la posizione fissa che i loro *values* occupano nel bastone di annotazione (cfr. § 6.2), manca la “ShN”, di fatto qui non usata.

MSF	1	pers=1	*posiz. 1*	8	degr=pos	*posiz. 4*
	2	pers=2		9	degr=comp	
	3	pers=3		10	degr=sup	
	4	gend=masc	*posiz. 2*	11	loc=const	*posiz. 5*
	5	gend=fem		12	loc=two	
	4;5	gend=c		13	loc=three	
	6	numb=sg	*posiz. 3*	14	loc=four	
	7	numb=pl		15	loc=five	
	6;7	numb=n		16	loc=six	
				17	loc=seven	
				18	loc=eight	
				19	loc=nine	

Tav.36: Le MSF: *feature declaration*

Le *HDF features* sono presentate nella forma di un *mapping* tra le tre notazioni interscambiabili, e cioè (da sinistra) “CdN” numerica, “ExN” (usata nella discussione precedente) e “ShN” (usata dal *query system*):

20	POS=n.type=com	n.c
21	POS=n.type=prop	n.p
26	POS=adj.type=qual	adj
30	POS=P-D.type=dem.infl=strg	pd.dem.s
31	POS=P-D.type=dem.infl=weak	pd.dem.w
32	POS=P-D.type=indf	pd.ind
33	POS=P-D.type=poss.infl=strg	pd.pos.s
34	POS=P-D.type=poss.infl=weak	pd.pos.w
35	POS=P-D.type=int	pd.int
36	POS=P-D.type=rel	pd.rel
37	POS=P-D.type=pers.infl=strg.case=nom	pd.per.s.n
38	POS=P-D.type=pers.infl=strg.case=obl	pd.per.s.o
39	POS=P-D.type=pers.infl=weak.case=obl	pd.per.w.o
40	POS=P-D.type=excl	pd.exc
41	POS=P-D.type=pers.infl=weak.case=nom	pd.per.w.n
45	POS=adv.type=general	adv.g
46	POS=adv.type=particle	adv.p
50	POS=conj.type=coord	con.c
51	POS=conj.type=subord	con.s

56	POS=adp.type=prepos	adp.pre
57	POS=adp.type=postpos	adp.post
60	POS=art.type=def	art.d
61	POS=art.type=indef	art.i
64	POS=num.type=card	num.c
65	POS=num.type=ord	num.o
68	POS=intj.type=general	intj
70	POS=punct.type=final	pun.fi
71	POS=punct.type=nonfinal	pun.nfi
75	POS=res.type=frgn	r.frg
76	POS=res.type=abbr	r.abb.
77	POS=res.type=formula	r.for
78	POS=res.type=epenth	r.epe
111	POS=v.type=mai.fin=fin.Vfm=ind.tns=pres	v.m.f.ind.pr
112	POS=v.type=mai.fin=fin.Vfm=ind.tns=ipf	v.m.f.ind.ipf
113	POS=v.type=mai.fin=fin.Vfm=ind.tns=past	v.m.f.ind.pt
114	POS=v.type=mai.fin=fin.Vfm=ind.tns=fut	v.m.f.ind.ft
115	POS=v.type=mai.fin=fin.Vfm=sub.tns=pres	v.m.f.sub.pr
116	POS=v.type=mai.fin=fin.Vfm=sub.tns=ipf	v.m.f.sub.ipf
117	POS=v.type=mai.fin=fin.Vfm=cond.tns=pres	v.m.f.cnd.pr
118	POS=v.type=mai.fin=fin.Vfm=impr.tns=pres	v.m.f.imp.pr
121	POS=v.type=mai.fin=no-fin.Vfm=inf	v.m.nf.inf.pr
122	POS=v.type=mai.fin=no-fin.Vfm=part.tns=pres	v.m.nf.par.pr
123	POS=v.type=mai.fin=no-fin.Vfm=part.tns=past	v.m.nf.par.pt
124	POS=v.type=mai.fin=no-fin.Vfm=ger.tns=pres	v.m.nf.ger.pr
211	POS=v.type=aux.fin=fin.Vfm=ind.tns=pres	v.a.f.ind.pr
212	POS=v.type=aux.fin=fin.Vfm=ind.tns=ipf	v.a.f.ind.ipf
213	POS=v.type=aux.fin=fin.Vfm=ind.tns=past	v.a.f.ind.pt
214	POS=v.type=aux.fin=fin.Vfm=ind.tns=fut	v.a.f.ind.ft
215	POS=v.type=aux.fin=fin.Vfm=sub.tns=pres	v.a.f.sub.pr
216	POS=v.type=aux.fin=fin.Vfm=sub.tns=ipf	v.a.f.sub.ipf
217	POS=v.type=aux.fin=fin.Vfm=cond.tns=pres	v.a.f.cnd.pr
218	POS=v.type=aux.fin=fin.Vfm=impr.tns=pres	v.a.f.imp.pr
221	POS=v.type=aux.fin=no-fin.Vfm=inf	v.a.nf.inf.pr
222	POS=v.type=aux.fin=no-fin.Vfm=part.tns=pres	v.a.nf.par.pr
223	POS=v.type=aux.fin=no-fin.Vfm=part.tns=past	v.a.nf.par.pt
224	POS=v.type=aux.fin=no-fin.Vfm=ger.tns=pres	v.a.nf.ger.pr
311	POS=v.type=mod.fin=fin.Vfm=ind.tns=pres	v.md.f.ind.pr
312	POS=v.type=mod.fin=fin.Vfm=ind.tns=ipf	v.md.f.ind.ipf
313	POS=v.type=mod.fin=fin.Vfm=ind.tns=past	v.md.f.ind.pt

314	POS=v.type=mod.fin=fin.Vfm=ind.tns=fut	v.md.f.ind.ft
315	POS=v.type=mod.fin=fin.Vfm=sub.tns=pres	v.md.f.sub.pr
316	POS=v.type=mod.fin=fin.Vfm=sub.tns=ipf	v.md.f.sub.ipf
317	POS=v.type=mod.fin=fin.Vfm=cond.tns=pres	v.md.f.cnd.pr
318	POS=v.type=mod.fin=fin.Vfm=impr.tns=pres	v.md.f.imp.pr
321	POS=v.type=mod.fin=no-fin.Vfm=inf	v.md.nf.inf.pr
322	POS=v.type=mod.fin=no-fin.Vfm=part.tns=pres	v.md.nf.par.pr
323	POS=v.type=mod.fin=no-fin.Vfm=part.tns=past	v.md.nf.par.pt
324	POS=v.type=mod.fin=no-fin.Vfm=ger.tns=pres	v.md.nf.ger.pr

Tav.37: Le HDF: *feature declaration*

6.2 IL BASTONE DI ANNOTAZIONE. L'annotazione complessiva che ogni "parola" (o meglio: token, cfr. qui Barbera - Corino - Onesti ¶3, § 1.3) si trova a ricevere nel testo, consistente nella associazione di lemma, nell'annotazione HDF ed in quella MSF, è quello che per comodità abbiamo deciso di chiamare, con espressione latamente tipografica, "bastone di annotazione" o più semplicemente "bastone".

Un "bastone vuoto", cioè una annotazione-tipo, ha la forma seguente:

*forma*_lem=*lemma*,HDF,MSF^1,MSF^2,MSF^3,MSF^4,MSF^5
es. torrai_lem=togliere,114,2,0,6,0,0

Tav.38: Il bastone di annotazione

Si noti peraltro che ogni bastone richiede sempre l'espressione di un valore (zero se nullo) per ogni posizione disponibile, per rendere possibile un riconoscimento anche posizionale dei codici.

6.3 LE ASSOCIAZIONI TRA HDF E MSF. Ogni HDF, come abbiamo visto nel prec. § 6.3, richiede l'obbligatoria specificazione di un *value* diverso da zero per un determinato *set* di MSF[52]; sfuggono a questo vincolo solo due HDF, 75 (res.frgn) e 76 (res.abbr), per le quali è possibile assegnare un *value* a qualsiasi MSF. La quinta MSF, loc, infine, a differenza delle precedenti quattro, può ricevere tanto "0" quanto un valore esplicito ("11-19") per qualsiasi HDF[53].

Le combinazioni obbligatorie[54] HDF+MSF sono dunque le seguenti:

[52] Nel nome, ad es., devono essere obbligatoriamente espressi genere e numero, nell'aggettivo genere, numero e grado, ecc.
[53] Almeno teoricamente: in pratica art, punct e adj non sembrano in italiano comprendere MW.
[54] Tra parentesi sono poste le due sopra accennate combinazioni ad espressione facoltativa.

HDF	***+ MSF***	***HDF***	***+ MSF***
20	gend,numb	113	pers,numb
21	gend,numb	114	pers,numb
26	gend,numb,degr	115	pers,numb
30	gend,numb	116	pers,numb
31		117	pers,numb
32	gend,numb	118	pers,numb
33	pers,gend,numb	121	
34	pers,gend,numb	122	gend,numb
35	gend,numb	123	gend,numb
36	gend,numb	124	
37	pers,gend,numb	211	pers,numb
38	pers,gend,numb	212	pers,numb
39	pers,gend,numb	213	pers,numb
40	gend,numb	214	pers,numb
41	pers,numb	215	pers,numb
45	degr	216	pers,numb
46		217	pers,numb
50		218	pers,numb
51		221	
56		222	gend,numb
57		223	gend,numb
60	gend,numb	224	
61	gend,numb	311	pers,numb
64	gend	312	pers,numb-
65	gend,numb	313	pers,numb
68		314	pers,numb
70		315	pers,numb
71		316	pers,numb
75	(pers,gend,numb,degr,loc)	317	pers,numb
76	(pers,gend,numb,degr,loc)	318	pers,numb
77		321	
78		322	gend,numb
111	pers,numb	323	gend,numb
112	pers,numb	324	

Tav.39: Le associazioni HDF+MSF nel CT tagset

In termini numerici le combinazioni sopra elencate si traducono nella seguente tabella, che praticamente esprime la struttura formale di tutti i bastoni (per l'espressione cfr. § 6.2) possibili nel nostro sistema di etichettatura:

20	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
21	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
26	0,4;5,6;7,8;9;10,0;11;12;13;14;15;16;17;18;19
30	4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
31	0,0,0,0,0
32	4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
33	1;2;3,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
34	1;2;3,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
35	4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
36	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
37	1;2;3,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
38	1;2;3,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
39	1;2;3,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
40	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
41	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
45	0,0,0,8;9;10,0;11;12;13;14;15;16;17;18;19
46	0,0,0,0,0
50	0,0,0,0,0;11;12;13;14;15;16;17;18;19
51	0,0,0,0,0;11;12;13;14;15;16;17;18;19
56	0,0,0,0,0;11;12;13;14;15;16;17;18;19
57	0,0,0,0,0;11;12;13;14;15;16;17;18;19
60	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
61	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
64	0,4;5,0,0,0;11;12;13;14;15;16;17;18;19
65	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
68	0,0,0,0,0;11;12;13;14;15;16;17;18;19
70	0,0,0,0,0
71	0,0,0,0,0
75	0;1;2;3,0;4;5,0;6;7,0;8;9;10,0;11;12;13;14;15;16;17;18;19
76	0;1;2;3,0;4;5,0;6;7,0;8;9;10,0;11;12;13;14;15;16;17;18;19
77	0,0,0,0,0
78	0,0,0,0,0
111	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
112	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
113	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
114	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
115	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
116	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
117	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
118	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
121	0,0,0,0,0;11;12;13;14;15;16;17;18;19
122	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
123	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
124	0,0,0,0,0;11;12;13;14;15;16;17;18;19
211	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
212	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
213	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
214	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
215	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19

216	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
217	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
218	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
221	0,0,0,0,0;11;12;13;14;15;16;17;18;19
222	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
223	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
224	0,0,0,0,0;11;12;13;14;15;16;17;18;19
311	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
312	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
313	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
314	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
315	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
316	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
317	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
318	1;2;3,0,6;7,0,0;11;12;13;14;15;16;17;18;19
321	0,0,0,0,0;11;12;13;14;15;16;17;18;19
322	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
323	0,4;5,6;7,0,0;11;12;13;14;15;16;17;18;19
324	0,0,0,0,0;11;12;13;14;15;16;17;18;19

Tav. 40: Le associazioni HDF+MSF: espressione numerica

7. UN ESEMPIO ANNOTATO: LA NOVELLA DI MASTRO TADDEO**.** Per concludere questa breve presentazione, voglio accludervi un piccolo esempio di un testo annotato tratto dal CT. Per esigenze di brevità devo presentarlo in notazione condensata, ma spero che ciò non crei troppi problemi[55]. Ho così scelto[56] la famosa novella di Mastro Taddeo ed il petronciano, che ha l'indubbio pregio di essere breve e divertente.

```
@Anonimo@@Novellino@@@Nar

%035

$0208$
Maestro       lem=maestro,20,0,4,6,0,0
Taddeo        lem=taddeo,21,0,4,6,0,0
,             lem=comma,71,0,0,0,0,0
leggendo      lem=leggere,124,0,0,0,0,0
a             lem=a,56,0,0,0,0,0
÷'            lem=il,60,0,4,7,0,0
suoi          lem=suo,33,3,4,7,0,0
scolari       lem=scolaio,20,0,4,7,0,0
in            lem=in,56,0,0,0,0,0
medicina      lem=medicina,20,0,5,6,0,0
venne         lem=venire,113,3,0,6,0,0
dinanzi       lem=dinanzi,45,0,0,0,8,0
a             lem=a,56,0,0,0,0,0
÷l            lem=il,60,0,4,6,0,0
maestro       lem=maestro,20,0,4,6,0,0
e             lem=e,50,0,0,0,0,0
disse         lem=dire,113,3,0,6,0,0
:             lem=colon,71,0,0,0,0,0

«             lem=guillemotleft,71,0,0,0,0,0
Maestro       lem=maestro,20,0,4,6,0,0
,             lem=comma,71,0,0,0,0,0
il            lem=il,60,0,4,6,0,0
cotale        lem=cotale,30,0,4;5,6,0,0
capitolo      lem=capitolo,20,0,4,6,0,0
```

[55] Poche ulteriori avvertenze: le fini di riga del testo in questa versione con *layout* verticale sono rappresentate dalla riga bianca; il numero dopo il simbolo del percento è quello della novella, mentre quello nel campo tra dollari fornisce la pagina; la riga con le chiocciole in testa fornisce gli identificativi di autore, titolo e genere; inoltre separati (cfr. § 5.2.7) con uno speciale codice (il *divide*, ASCII `Alt+246` = ANSI `Alt+0247`) ed annotati individualmente sono gli elementi in clisi grafica.

Sono quegli elementi che nella nostra procedura abbiamo scelto di designare come "grafoclitici", comprendendovi oltre ai clitici "veri" quando scritti unitamente alla parola di appoggio (quelli cioè di *dimmelo*, ma non quelli di *me lo dici*) anche gli articoli delle preposizioni articolate, ed in genere tutti gli elementi in analoghe condizioni grafiche (ad es. la "postposizione" in *meco*, ecc.).

[56] Per ragioni di spazio il testo ha dovuto essere molto rimpicciolito.

,	lem=comma,71,0,0,0,0,0
trovò	lem=trovare/-si/,113,3,0,6,0,0
che	lem=che,51,0,0,0,0,0
,	lem=comma,71,0,0,0,0,0
chi	lem=chi,36,0,4;5,6;7,0,0
continuo	lem=continuo,45,0,0,0,8,0
mangiasse	lem=mangiare,116,3,0,6,0,0
nove	lem=nove,64,0,4;5,0,0,0
dì	lem=dì,20,0,4,6,0,0
di	lem=di,56,0,0,0,0,0
petronciani	lem=petronciano,20,0,4,7,0,0
,	lem=comma,71,0,0,0,0,0
che	lem=che,51,0,0,0,0,0
diverrebbe	lem=divenire,117,3,0,6,0,0
matto	lem=matto,26,0,4,6,8,0
;	lem=semicolon,71,0,0,0,0,0
e	lem=e,50,0,0,0,0,0
provava	lem=provare,112,3,0,6,0,0
÷lo	lem=lo,39,3,4,6,0,0
secondo	lem=secondo,56,0,0,0,0,0
fisica	lem=fisica,20,0,5,6,0,0
.	lem=stop,70,0,0,0,0,0
Un	lem=uno,61,0,4,6,0,0
suo	lem=suo,33,3,4,6,0,0
scolaro	lem=scolaio,20,0,4,6,0,0
,	lem=comma,71,0,0,0,0,0
udendo	lem=udire,124,0,0,0,0,0
quel	lem=quello,30,0,4,6,0,0
capitolo	lem=capitolo,20,0,4,6,0,0
,	lem=comma,71,0,0,0,0,0
propuose	lem=proporre/-si/,113,3,0,6,0,0
÷si	lem=si,39,3,4;5,6;7,0,0
di	lem=di,51,0,0,0,0,0
voler	lem=volere/-si/,321,0,0,0,0,0
÷lo	lem=lo,39,3,4,6,0,0
provare	lem=provare,121,0,0,0,0,0
:	lem=colon,71,0,0,0,0,0
prese	lem=prendere,113,3,0,6,0,0
a	lem=a,51,0,0,0,0,0
mangiare	lem=mangiare,121,0,0,0,0,0
de	lem=di,56,0,0,0,0,0
÷'	lem=il,60,0,4,7,0,0
petronciani	lem=petronciano,20,0,4,7,0,0
,	lem=comma,71,0,0,0,0,0
et	lem=e,50,0,0,0,0,0
in	lem=in,56,0,0,0,0,0
capo	lem=capo,20,0,4,6,0,0
de	lem=di,56,0,0,0,0,0
÷'	lem=il,60,0,4,7,0,0
nove	lem=nove,64,0,4;5,0,0,0
dì	lem=dì,20,0,4,6,0,0
che	lem=che,36,0,4;5,6;7,0,0
leggeste	lem=leggere,113,2,0,6,0,0
non	lem=non,45,0,0,0,8,0
è	lem=essere,211,3,0,6,0,0
vero	lem=vero,26,0,4,6,8,0
,	lem=comma,71,0,0,0,0,0
però	lem=però,51,0,0,0,0,0
ch'	lem=che,51,0,0,0,0,0
io	lem=io,37,1,4;5,6,0,0
l'	lem=lo,39,3,4,6,0,0
ho	lem=avere,211,1,0,6,0,0
÷e	lem=÷e,78,0,0,0,0,0
provato	lem=provare,123,0,4,6,0,0
,	lem=comma,71,0,0,0,0,0
e	lem=e,50,0,0,0,0,0
non	lem=non,45,0,0,0,8,0
sono	lem=essere,211,3,0,7,0,0
matto	lem=matto,26,0,4,6,8,0
»	lem=guillemotright,71,0,0,0,0,0
:	lem=colon,71,0,0,0,0,0
e	lem=e,50,0,0,0,0,0
pure	lem=pure,45,0,0,0,8,0
alza	lem=alzare,111,3,0,6,0,0
÷si	lem=si,39,3,4;5,6;7,0,0
e	lem=e,50,0,0,0,0,0
mostro	lem=mostrare,113,3,0,6,0,0
÷lli	lem=gli,39,3,4,6;7,0,0
il	lem=il,60,0,4,6,0,0
culo	lem=culo,20,0,4,6,0,0
.	lem=stop,70,0,0,0,0,0
0209	
«	lem=guillemotleft,71,0,0,0,0,0
Iscrivete	lem=scrivere,118,2,0,7,0,0
»	lem=guillemotright,71,0,0,0,0,0
disse	lem=dire,113,3,0,6,0,0
il	lem=il,60,0,4,6,0,0
maestro	lem=maestro,20,0,4,6,0,0
«	lem=guillemotleft,71,0,0,0,0,0
che	lem=che,51,0,0,0,0,0
provato	lem=provare,123,0,4,6,0,0
è	lem=essere,211,3,0,6,0,0
;	lem=semicolon,71,0,0,0,0,0
e	lem=e,50,0,0,0,0,0
faccia	lem=fare/-si/,115,1;2;3,0,6,0,0
÷se	lem=si,39,3,4;5,6;7,0,0
÷ne	lem=ne,31,0,0,0,0,0
nuova	lem=nuovo,26,0,5,6,8,0
chiosa	lem=chiosa,20,0,5,6,0,0
»	lem=guillemotright,71,0,0,0,0,0
.	lem=stop,70,0,0,0,0,0

Tav. 41: La novella di Mastro Taddeo POS-taggata

BIBLIOGRAFIA.

AA. VV.
2004 *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, ELDA, 2004.

ALLEGRANZA - MAZZINI
2000 Valerio Allegranza - Giampaolo Mazzini, *Linguistica generativa e grammatiche a unificazione*, Torino, Paravia, 2000 "Scriptorium. Sapere linguistico e pratica dell'italiano".

ARMSTRONG
1994 *Using Large Corpora*, edited by Susan Armstrongs, Cambridge (Mass.) - London (En.), The MIT Pres, 1994 "A Bradford Book", "ACL-MIT Press Series in Computational Linguistics" [= "Compuational Linguistics" XIX (1993)[1-2]].

ATWELL - SOUTER 1993 → SOUTER - ATWELL 1993

BARBERA
2001 Manuel Barbera, *From EAGLES to CT Tagging: a Case for Re-usability of Resources*, in RAYSON et alii 2001, pp. 40-44.
2000/2002 Manuel Barbera, *Pronomi e determinanti nell'annotazione dell'italiano antico. La POS "PD" del Corpus Taurinense*, in BAUER - GOEBL 2002, pp. 35-52.
2000/2006 Manuel Barbera, *CT Specification Guide*, HTML page, 29 August 2000, nel sito ospitato dall'IMS di Stuttgart dal titolo *WWW access to the corpus Corpus Taurinense (XIIIth century Italian)*: `http://www.ims.uni-stuttgart.de/projekte/CQPDemos/italant/` e poi nel sito di corpora.unito.it `http://www.corpora.unito.it/italant/posinfo.html`. La versione più recente (2006) è però sempre quella disponibile alla pagina `http://www.bmanuel.org/projects/ct-posinfo.htm`.
¶ iiij Manuel Barbera, *La resa dei forestierismi in italiano. Breve nota ortografica*, in questo volume, pp. xv-xvj.
¶ 1 Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in questo volume, pp. 3-20.
¶ 23 Manuel Barbera, *Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra* guidelines *e prolegomeni*., in questo volume, pp. 373-388.

BARBERA - CORINO - ONESTI
¶ 3 Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in questo volume, pp. 25-88.

BARBERA - MARELLO
1999/2001 Manuel Barbera - Carla Marello, *L'annotazione morfosintattica del Padua Corpus: strategie adottate e* problemi *di acquisizione*, comunicazione al convegno *Italiano antico e corpora elettronici*, Padova, 19-20 febbraio 1999, poi in "Révue romane" XXXVI (2001)[1] 3-20.
2000 Manuel Barbera - Carla Marello, *Les lexies complexes et leur annotation morphosyntactique dans le Corpus Taurinense*, intervento al convegno AFLA 2000, Paris, 6-8 luglio 2000, poi in "Révue française de linguistique appliquée" V (2000)[2] "Dossier. Diversité du traitement automatique des langues" pp. 57-70.
2000/2003 Manuel Barbera - Carla Marello, *Corpus Taurinense: italiano antico annotato in modo nuovo*, in MARASCHIO - POGGI SALANI 2003, pp. 685-693.

2003 *i.s.* Manuel Barbera - Carla Marello, *Corpo a corpo con l'inglese della* corpus linguistics*, anzi, della linguistica dei corpora*, in *Atti del Convegno Internazionale Lingua italiana e scienze , Firenze, Accademia della Crusca 6-8 febbraio 2003*, in corso di stampa.

BARONI et alii

2004 Marco Baroni - Silvia Bernardini - Federica Comastri - Lorenzo Piccioni - Alessandra Volpi - Guy Aston - Marco Mazzoleni, *Introducing the* La Repubblica *Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*, in AA. VV. 2004, pp. 1771-1774, disponibile online alla pagina `http://www.form.unitn.it/~baroni/publications/lrec2004/rep_lrec_2004.pdf`.

BAUER - GOEBL

2002 *Parallela IX. Testo - variazione - informatica | Text - Variation - Informatik. Atti del IX Incontro italo-austriaco dei linguisti (Salisburgo, 1-4 novembre 2000) | Akten des IX Österreichisch-italienischen Linguistentreffens (Salzburg, 1.-4. November 2000)*, a cura di | hrsg. von Roland Bauer - Hans Goebl, Wilhelmsfeld, Gottfied Egert, 2002 "Pro Lingua" 35

BEGGIATO - MARINETTI - MARRONI

2002 Fabrizio Beggiato - Sabina Marinetti - Sergio Marroni, *AMIA (Analizzatore Morfosintattico dell'Italiano Antico)*, in "La comunicazione" XIII (2002) 149-150; disponibile online alla pagina `http://www.iscom.gov.it/documenti/files/rivista/2002_149.pdf`. [numero speciale: *Atti della conferenza TIPI: Tecnologie Informatiche nella Processazione della Lingua Italiana*; versione online: `http://www.iscom.gov.it/contenuti. asp?ID=140&sID=24&xsID=81`]

BELTRAMI

1983-... *Tesoro della lingua italiana delle origini*, diretto da Pietro Beltrami, Firenze, CNR - Centro di studi Opera del Vocabolario Italiano, 1983-..., disponibile su `http://www.csovi.fi.cnr.it/`.

BRESNAN

1982 *The Mental Representation of Grammatical Relations*, edited by Joan Bresnan, Cambridge (Mass.), MIT Press, 1982.

CABRÉ - MOREL - TORNER - VIVALDI - YZAGUIRRE

1998 Maria Teresa Cabré - Jordi Morel - Sergi Torner - Jordi Vivaldi - Lluis de Yzaguirre, *El corpus de l'IULA: etiquetaris*, Barcelona, Universitat Pompeu Fabra. Institut Universitari de Lingüistica Aplicada, 1998 "Sèrie Informes" 18; disponibile anche online con la sigla IULA/INF018/98 alla pagina `http://www.iula.upf.es/paps1ca.htm`.

CARNAP

1937/1934 Rudolf Carnap, *The Logical Syntax of Language*, English translation by Amethe Smeaton Countess von Zeppelin, London: Routledge & Kegan Paul, 1937 [1967[7]; edizione originale *Logische Syntax der Sprache*, Wien 1934].

1974/1963 Rudolf Carnap, *Autobiografia intellettuale*, in *La filosofia di Rudolf Carnap*, a cura di Paul Arthur Schilpp, trad. di Maria Grazia Cristofaro Sandrini, Milano, il Saggiatore ("Biblioteca di filosofia e metodo scientifico"), 1974 pp. 1-85 e 997-998 [edizione originale "Intellectual Autobiography", in *The Philosophy of Rudolf Carnap*, edited by P[aul] A[rthur] Schilpp, La Salle (Illinois), The Library of Living Philosophers, 1963].

CASTELLANI
1965 Arrigo Castellani, *Pisano e lucchese*, in "Studi linguistici italiani" V (1965) 97-135; poi in CASTELLANI 1980, vol. I, pp. 283-326.
1980 Arrigo Castellani, *Saggi di linguistica e filologia italiana e romanza (1946-1976)*, Roma, Salerno Editrice, 1980, voll. 1-3.

CHRIST - SCHULZE
1996 Oliver Christ - Bruno Maximilian Schulze, *CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora*, in FELDWEG - HINRICHS 1996; disponibile online alla pagina `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz`.

DÖRRE - DORNA
1993 Jochen Dörre - Michael Dorna, *CUF - A Formalism for Linguistic Knowledge Representation*, Deliverable R.1.2A, DYANA 2. Postscript version (R1_2_A-Dorre1.ps), disponibile online alla pagina `http://www.essex.ac.uk/linguistics/clmt/papers/cuf/`.

EAGLES (ELM-DE, ELM-EN, ELM-FR, ELM-IT, MORPHSYN) → TEUFEL - STÖCKERT 1996, TEUFEL 1996, REKOVSKI 1996, MONACHINI 1996, MONACHINI - CALZOLARI 1996.

ELM-DE → TEUFEL - STÖCKERT 1996.

ELM-EN → TEUFEL 1996.

ELM-FR → REKOVSKI 1996.

ELM-IT → MONACHINI 1996.

FELDWEG - HINRICHS
1996 *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, herausgegeben von Helmut Feldweg und Erhard W. Hinrichs, Tübingen, Max Niemeyer Verlag, 1996 "Lexicographica. Series maior" 73.

FELDWEG - KIBIGER - THIELEN
1995 Helmut Feldweg - Ralf Kibiger - Christine Thielen, *Zum Sprachgebrauch in deutschen Newsgruppen*, in "Osnabrücker Beiträge zur Sprachtheorie" L (1995) 143-154, disponibile anche online `http://www.sfs.uni-tuebingen.de/Elwis/news.ps`.

GARSIDE - LEECH - MCENERY
1997 *Corpus Annotation. Linguistic Information from Computer Text Corpora*, edited by Roger Garside, Geoffrey Leech and Anthony McEnery, London - New York, Longman, 1997.

GARSIDE - LEECH - SAMPSON
1987 *The Computational Analysis of English: a Corpus-based Approach*, edited by Roger Garside, Geoffrey Leech and Geoffrey Sampson, London - New York, Longman, 1987.

GGIC I → RENZI - SALVI et alii 1988; II → RENZI - SALVI et alii 1991; III → RENZI - SALVI et alii 1995.

GRAFFI
1991 Giorgio Graffi, *Concetti 'ingenui' e concetti 'teorici' in sintassi*, in "Lingua e stile" XXVI (1991) 347-363.
1994 Giorgio Graffi, *Sintassi*, Bologna, il Mulino, 1994 "Strumenti. Le strutture del linguaggio" [4].

GREENBAUM
1993 Sidney Greenbaum, *The Tagset for the International Corpus of English*, in SOUTER - ATWELL 1993, pp. 11-24.

ItalAnt → RENZI - SALVI *i.s*.

HEID
1998 Ulrich Heid, *Annotazione morfosintattica di corpora ed estrazione di informazioni linguistiche*, relazione al convegno *Annotazione morfosintattica di corpora e costruzione di banche di dati linguistici. Torino, 26-XI-1998*, inedita.

IORIO-FILI
1997 Domenico Iorio-Fili, *Un nuovo software lessicografico: GATTO*, in “Opera del Vocabolario italiano. Bollettino” II (1997) 259-270.

KAPLAN - BRESNAN
1982 Roland M. Kaplan - Joan Bresnan, *Lexical-Functional Grammar: a Formal System for Grammatical Representation*, in BRESNAN 1982, pp. 173-381.

KARLSSON et alii
1995 *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*, edited by Fred Karlsson, Atro Voutilainen, Juha Heikkilä and Arto Anttila, Berlin and New York, Mouton de Gruyter, 1995 “Natural Language Processing” 4.

KERMES - EVERT
2002 Hannah Kermes - Stefan Evert, *YAC -- A Recursive Chunker for Unrestricted German Text*, in RODRIGUEZ - SUAREZ ARAUJO 2002, volume V, pp. 1805-1812; disponibile online alla pagina `http://www.ims.uni-stuttgart.de/~kermes/publications.shtml`.

KÖNIG
1996 Esther König, *Introduction to Categorial Grammars*, Stuttgart, IMS, May 1996. Online alla pagina `http://www.ims.uni-stuttgart.de/projekte/cuf/LexGram/LexGram.html`.

LEECH
1997 Geoffrey Leech, *Introducing Corpus Annotation*, in GARSIDE - LEECH - MCENERY 1997, pp. 1-18.
1997a Geoffrey Leech, *Grammatical Tagging*, in GARSIDE - LEECH - MCENERY, pp. 19-33.

LEECH - WILSON
1999 Geoffrey Leech - Andrew Wilson, *Standards for Tagsets*, in VAN HALTEREN 1999, pp. 55-80.

LEOPARDI
1817-27/1991 Giacomo Leopardi, *Zibaldone di pensieri*, edizione critica e annotata a cura di Giuseppe Pacella, Milano, Garzanti, 1991 “I libri della spiga”.

MARASCHIO - POGGI SALANI
2003 *Italia linguistica anno Mille - Italia linguistica anno Duemila. Atti del XXIV Congresso internazionale di studi della Società di linguistica italiana (SLI), Firenze 19-21 ottobre 2000*, a cura di Nicoletta Maraschio e Teresa Poggi Salani, Roma Bulzoni, 2003.

MARCUS - SANTORINI - MARCINKIEVICZ
1994 Mitchell P. Marcus - Beatrice Santorini - Mary Ann Marcinkievicz, *Building a Large Annotated Corpus of English: The Penn Treebank*, in ARMSTRONG 1994, pp. 273-290. Disponibile online dalla homepage del PennTreebank al link `ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz`.

MONACHINI
1996 Monica Monachini, *ELM-IT: EAGLES Specifications for Italian Morphosyntax - Lexicon Specifications and Classification Guidelines*, Pisa, EAGLES Document EAG-CLWG-ELM-IT/F, May 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`.

MONACHINI - CALZOLARI
1996 Monica Monachini - Nicoletta Calzolari, S*ynopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Application to European Languages*, Pisa, EAGLES Document EAG-CLWG-MORPHSYN/R, May 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`.
1999 Monachini, Monica - Calzolari, Nicoletta, *Standardization in the Lexicon*, in VAN HALTEREN 1999, pp. 149-174.

MORPHSYN → MONACHINI - CALZOLARI 1996.

PACELLA 1991 → LEOPARDI 1817-27/1991

POLLARD - SAG
1987 Carl Pollard - Ivan A. Sag, *Information-Based Syntax and Semantics*, Stanford, Stanford University Centre for the study of language and information, 1987 "CSLI lecture notes" 13.

RAYSON et alii
2001 *Proceedings of the Corpus Linguistics 2001 Conference. Lancaster University 29 March - 2 April 2001*, edited by Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja, Lancaster, University Center for Computer Corpus Research on Language, 2001 "UCREL Technical Paper" 13.

REKOWSKI
1996 Ursula von Rekowski, *Specifications for French Morphosyntax - (ELM-FR)*, Paris, EAGLES Document EAG-CLWG-ELM-FR/F, 31st Aug. 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`

RENZI
1998 *ITALANT: per una Grammatica dell'Italiano Antico*, a cura di Lorenzo Renzi, Padova, Centro Stampa di Palazzo Maldura, 1998.
1998a Lorenzo Renzi, *Perché una grammatica dell'italiano antico: una presentazione*, in Renzi 1998, pp. 21-32.

RENZI - SALVI et alii
1988 *Grande grammatica italiana di consultazione*. Volume I, *La frase. I sintagmi nominale e preposizionale*, a cura di Lorenzo Renzi, Bologna, il Mulino, 1988.
1991 *Grande grammatica italiana di consultazione*. Volume II, *I sintagmi verbale, aggettivale, avverbiale. La* subordinazione, a cura di Lorenzo Renzi e Giampaolo Salvi. Bologna, il Mulino, 1991.

1995 *Grande grammatica italiana di consultazione*. Volume III, *Tipi di frase, deissi, formazione delle parole*, a cura di Lorenzo Renzi, Giampaolo Salvi e Anna Cardinaletti. Bologna, il Mulino, 1995.

i.s. *ItalAnt. Grammatica dell'italiano antico*, a cura di Lorenzo Renzi e Giampaolo Salvi, Bologna, il Mulino, in corso di stampa.

RODRIGUEZ - SUAREZ ARAUJO

2002 *Proceedings of the Third International Conference on Language Resources and Evaluation*, edited by Manuel Gonzalez Rodriguez and Carmen Paz Suarez Araujo, 2002.

ROHLFS

1966-69 *Grammatica storica della lingua italiana e dei suoi dialetti*, Vol. I. *Fonetica*. Traduzione di Salvatore Persichino, Vol. II. *Morfologia*. Traduzione di Temistocle Franceschi, Vol. III. *Sintassi e formazione delle parole*.Traduzioni di Temistocle Franceschi e Maria Ciagagli Franceschi, Torino, Einaudi, risp. 1966, 1968 e 1969 "Piccola Biblioteca Einaudi" 148, 149 e 150.

SANTORINI

1990/1 Beatrice Santorini, *Part-of-speech Tagging Guidelines for the Penn Treebank Project*, Technical report MS-CIS-90-47, University of Pennsylvania - Department of Computer and Information Science, 1990. *3rd Revision, 2nd Printing, June 1990* è disponibile online dalla homepage del PennTreebank `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`; la *Rev. 1991 March 15* è disponibile dalla homepage del Treetagger al link `http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps`.

SCHILLER - STÖCKERT - TEUFEL - THIELEN

1999 Anne Schiller - Simone Teufel - Christine Stöckert - Christine Thielen, *Guidelines für das Tagging Deutscher Textkorpora mit STTS. (Kleines und großes Tagset)*, Technical report, IMS and SfS, disponibile online alla pagina `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.ps.gz`

SCHILLER - TEUFEL - THIELEN

1995 Anne Schiller - Simone Teufel - ChristineThielen, *Guidelines für das Tagging Deutscher Textkorpora mit STTS*, IMS and SfS, Draft 26 September 1995, disponibile online a `http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz`

SCHMID

1994 Helmut Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, paper presented at the *International Conference on New Methods in Language Processing*, Manchester (UK), 1994; versione revisionata PS/PDF online sul sito dell'IMS Stuttgart: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

SOUTER - ATWELL

1993 *Corpus-based Computational Linguistics*, edited by Clive Souter and Eric Atwell, Amsterdam - Atalanta, Rodopi, 1993 "Language and Computers: Studies in Practical Linguistics" 9.

TEUFEL
1996 Simone Teufel, *ELM-EN. EAGLES Specifications for English Morphosyntax. Draft Version*, Stuttgart, EAGLES Document, July, 31 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`

TEUFEL - STÖCKERT
1996 Simone Teufel - Christine Stöckert, *ELM-DE. EAGLES Specification for German Morphosyntax. Lexicon Specification and Classification Guidelines*, Stuttgart, EAGLES Document EAG-CLWG-ELM-DE/F, März 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`

VAN HALTEREN
1999 *Syntactic Wordclass Tagging*, edited by Hans van Halteren, Dordrecht - Boston - London, Kluver Academic Publishers, 1999 "Text, Speech and Language Technology" 9.

CORPORA, STRUMENTI E SITI DI RIFERIMENTO.

AMIA `http://www.iscom.gov.it/documenti/files/rivista/2002_149.pdf`

Brown Corpus `http://en.wikipedia.org/wiki/Brown_Corpus`
`http://ota.ahds.ac.uk/` (*search*)

CG2 `http://www.ling.helsinki.fi/~tapanain/cg/index.html`

CiBIT `http://cibit.humnet.unipi.it/index_ra.htm`

Corpus Taurinense `http://www.bmanuel.org/projects/ct-HOME.html`

CT → Corpus Taurinense

CUF `http://www.ims.uni-stuttgart.de/projekte/cuf/`

CWB `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`

EAGLES `http://www.ilc.cnr.it/EAGLES96/home.html`

ICAME `http://icame.uib.no/`

IMS Stuttgart `http://www.ims.uni-stuttgart.de/ims-home. html.en`

ISLE `http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm`

ItalAnt `http://geocities.com/gpsalvi/konyv/`

IULA Corpora `http://www.iula.upf.es/corpus/corpusuk.htm`

LLC `http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM` (corpus disponibile da ICAME)
`http://ota.ahds.ac.uk/` (*search*)
`ftp://ftp.cogsci.ed.ac.uk/pub/corpus-LLC/`

LOB Corpus `http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html#lob`
`http://ota.ahds.ac.uk/` (*search*)

OVI db testuale	http://ovisun198.ovi.cnr.it/italnet/OVI/
Penn Treebank	http://www.cis.upenn.edu/~treebank/home.html
PPCME	http://www.ling.upenn.edu/hist-corpora/
Stein homepage	http://www.uni-stuttgart.de/lingrom/stein/
STTS	http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html
TBPCHP	http://www.ime.usp.br/~tycho/corpus/files/index.html
Tree Tagger	http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
UCREL	http://www.comp.lancs.ac.uk/computing/research/ucrel/

MARCO TOMATIS
(Torino)

9. La disambiguazione del Corpus Taurinense. *Problemi teorici e pratici.*

0. INTRODUZIONE. Una delle prerogative distintive del Corpus Taurinense consiste nell'aver subito un consistente processo di disambiguazione come passo successivo al *POS-tagging* (come già accennato in Barbera ¶ 1, *supra*, § 2.2.1). Tale processo si è reso indispensabile al fine di garantire l'assegnazione univoca delle varie classi grammaticali ai diversi token costituenti l'intero testo. Nel seguente articolo si illustreranno, in modo preliminare, i problemi teorici e tecnici con cui è stato necessario confrontarsi per la disambiguazione del Corpus Taurinense e, più nel dettaglio, le procedure e le soluzioni computazionali successivamente adottate.

0.1 SISTEMI DI DISAMBIGUAZIONE: UNA PANORAMICA GENERALE. Per una trattazione il più possibile chiara del tema in questione, si rivela indispensabile fornire una descrizione di massima dei vari significati che il termine "disambiguazione" può veicolare all'interno del paradigma della *corpus linguistics*. È bene precisare, tuttavia, che sebbene tale termine esprima chiaramente il concetto di eliminazione o riduzione del grado di ambiguità posseduto da un determinato elemento presente all'interno di un sistema complesso, quale quello lessicale, la trattazione prenderà in esame unicamente il livello di analisi di natura testuale, tralasciando le questioni relative alla gestione di informazioni di tipo sonoro. Fatta questa premessa, è bene chiarire subito che con "disambiguazione" è possibile fare riferimento a due generi di problemi differenti e ben distinti tra loro. La disambiguazione di un dato elemento testuale, infatti, può riferirsi sia alla definizione univoca delle caratteristiche semantiche che tale elemento possiede, naturalmente in stretta relazione con il contesto in cui si trova inserito, sia alla definizione univoca delle sue caratteristiche in termini di categoria grammaticale di appartenenza (indicata anche con POS, ossia *part of speech*). Sebbene l'elaborazione computazionale della semantica dei vari token costituenti il testo sia un settore di ricerca molto complesso e in piena evoluzione, che richiede l'uso di strumenti appropriati quali ontologie e reti semantiche, è utile precisare che per quanto concerne il processo di disambiguazione lessicale progettato per il Corpus Taurinense, tale elaborazione si limita alla seconda delle accezioni di cui sopra, ossia al livello delle categorie morfosintattiche. A tale proposito bisogna ricordare che a differenza della disambiguazione testuale di natura semantica, obbligatoriamente vincolata all'analisi del contesto specifico, il processo di disambiguazione morfosintattico può essere elaborato secondo modelli computazionali sia di tipo *context sensitive* (sensibili al contesto), sia *context free* (svincolati dal contesto). Al riguardo è utile precisare che questi ultimi, essendo unicamente legati alla natura morfosintattica dei token circostanti, ma non alla loro forma lessicale, risultano intrinsecamente dotati di maggiore potenza e flessibilità rispetto ai primi.

In merito agli strumenti necessari per ottenere la disambiguazione lessicale a livello di categoria grammaticale, l'operatore può scegliere di optare per due diverse soluzioni alternative. La prima di esse, utilizzata nella maggior parte dei casi, prevede l'adozione di sistemi stocastici basati su Modelli Markoviani Nascosti (HMM, *Hidden Markov Models*), strumenti molto potenti e versatili che coniugano in una sola fase il processo di assegnazione delle etichette morfosintattiche e la conseguente disambiguazione. Nonostante le già citate doti di robustezza e flessibilità, è importante tuttavia sottolineare la necessità di tali sistemi di avvalersi di un corpus di dimensioni ridotte, precedentemente annotato, da cui poter trarre le informazioni utili per svolgere l'elaborazione statistica. Tale procedura, denominata *training* (ossia allenamento) del tagger,

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp.169-181.

risulta a tutti gli effetti indispensabile, tanto da risultare un elemento fondamentale per la valutazione delle prestazioni del sistema.

Per contro, la seconda delle soluzioni possibili prevede di limitare la fase di etichettatura alla semplice assegnazione delle varie POS a tutti i token che compongono il testo, proseguendo successivamente l'elaborazione con la fase di disambiguazione vera e propria, solitamente costituita da un motore basato su regole linguistiche di tipo *context free* o, meno preferibilmente, *context sensitive*. Naturalmente, a differenza dei sistemi basati su modelli statistici, discussi in precedenza, nel caso in questione sarà necessario l'utilizzo di due processi distinti e nettamente separati anche sotto il profilo dell'elaborazione computazionale.

Per quanto riguarda un bilancio sommario dei pregi e dei difetti di entrambi i sistemi, risulta evidente che il processo stocastico, dato il ridotto numero di fasi di elaborazione, risulti caratterizzato da notevoli vantaggi in termini di velocità di esecuzione e leggerezza computazionale. A tali caratteristiche si somma l'interessante capacità di poter assegnare POS univoche anche laddove la parola specifica risulti totalmente ignota al lessico di riferimento del sistema (es. neologismi). Per contro, la necessità di disporre di un *training corpus* preventivo ne riduce drasticamente la capacità di utilizzo al di fuori dei canoni linguistici già noti e consolidati. I modelli a regole, per contro, pur essendo svincolati dalla necessità di disporre di un corpus già annotato, con il conseguente vantaggio di poter essere applicati anche su corpora di lingue mai precedentemente etichettate, risultano di fatto assai più pesanti in termini di elaborazione computazionale. Inoltre, la necessità di un lungo lavoro di sviluppo di regole linguistiche sulla base di un formalismo ben preciso e definito, rende il sistema maggiormente costoso, nonché molto meno agevole da gestire e mantenere.

Per quanto riguarda il CT, risulta evidente che, nonostante i difetti emersi, volendo trattare una lingua computazionalmente vergine come l'italiano del '200, l'unica soluzione possibile fosse lo sviluppo di un sistema articolato di regole capace di coprire l'intera gamma di possibili varianti e anomalie linguistiche presenti all'interno del corpus.

0.2 PREMESSE METODOLOGICHE. Lo sviluppo di un sistema di disambiguazione contestuale del *Corpus Taurinense* si è presentato fin dai primi momenti come un'opera di non banale complessità. Di diversa natura, infatti, sono i problemi che deve affrontare la persona che si accinge a compiere tale opera: il primo, e più evidente, consiste nella natura del corpus stesso. Trattandosi di una lingua antica, infatti, è necessario l'ausilio di una persona dotata di un buon bagaglio filologico al fine di ottenere una corretta interpretazione del testo e, conseguentemente, una corretta gestione delle diverse problematiche linguistiche che possono presentarsi durante lo svolgimento del lavoro. Il secondo tipo di difficoltà, di natura più eminentemente pratica, consiste nella necessità di scegliere un formalismo od un linguaggio di programmazione che risulti il più adeguato possibile allo scopo che si vuole portare a termine, senza tuttavia introdurre un eccessivo livello di complessità computazionale o difficoltà realizzativa, elementi questi che potrebbero distogliere energie al più importante problema dell'effettiva formulazione della grammatica di disambiguazione.

Se il problema di natura filologica ha visto una soluzione piuttosto agevole grazie al prezioso contributo di Manuel Barbera, la decisione in merito alla tipologia del sistema computazionale da adottare ha richiesto uno sforzo valutativo più intenso. L'elaborazione elettronica delle lingue naturali (NLP - *Natural Language Processing*) dispone di numerosi strumenti informatici, perlopiù linguaggi di programmazione, caratterizzati da peculiarità che permettono di conseguire i risultati desiderati nella maniera più agevole possibile. Pertanto, se un efficiente sistema di analisi morfologica può essere realizzato mediante un automa a stati finiti non deterministici sviluppato in un linguaggio multipiattaforma quale il Java, un sistema di analisi sintattica (*parser*) può essere altrettanto agevolmente prodotto mediante l'uso di un linguaggio dichiarativo basato sulla logica matematica quale il Prolog (termine composto dalla sigla "*Programming in*

Logic”). In seguito a numerose valutazioni tecnico-pragmatiche si è deciso di implementare la grammatica di disambiguazione in una struttura di programma basata sul linguaggio AWK. Tale scelta, forse criticabile per alcuni aspetti di natura più marcatamente informatica relativi a valutazioni di velocità ed efficienza computazionale, ha avuto tuttavia il merito di fornire al sistema di regole (che ricordiamo essere in maggior numero di tipo *context sensitive*, ossia strettamente legate al contesto in cui operano) una struttura estremamente flessibile, leggera, versatile e facilmente adattabile ad ulteriori aggiunte o modifiche.

1. ARCHITETTURA DEL SISTEMA DI DISAMBIGUAZIONE. Data la natura tipicamente procedurale del linguaggio adottato, il sistema di disambiguazione possiede una struttura generale costituita da una serie di moduli indipendenti, operanti secondo una ben precisa gerarchia sequenziale. Attualmente il sistema si compone di sei moduli di disambiguazione e due moduli di formattazione del testo, la cui funzione verrà discussa più avanti. Poiché soltanto il primo dei sei moduli opera su una copia opportunamente formattata del testo etichettato originale, mentre ogni modulo successivo agisce sul testo generato dall'elaborazione del modulo precedente, ecco che l'organizzazione del sistema in una ben precisa gerarchia d'intervento si rivela una soluzione indispensabile. Tale configurazione, infatti, consente di frazionare e distribuire le operazioni di disambiguazione in vari livelli distinti, secondo una disposizione gerarchica che è funzione della rilevanza linguistica e computazionale delle varie POS[1] (*part of speech*) da trattare. Non risulta casuale, quindi, che il modulo iniziale sia composto unicamente dalle regole atte a trattare le forme caratterizzate da ambiguità nome / verbo (es. “fatto”), mentre il successivo comprenda le forme nome / aggettivo non disambiguabili da regole generali.

La struttura interna dei singoli moduli risulta piuttosto semplice: ogni modulo è costituito da una serie di regole a mutua esclusione che agiscono sul testo etichettato come una sorta di *filtro passivo*. L'intero processo di disambiguazione, infatti, si limita ad eliminare le voci di transcategorizzazione non pertinenti semplicemente assegnando, previa selezione, l'elemento morfosintattico più corretto all'interno di ogni token caratterizzato da ambiguità.

Esistono due diversi tipi di ambiguità che il sistema qui descritto è in grado di riconoscere e correggere. Definiti rispettivamente con i termini di *ambiguità interna* ed *ambiguità esterna*, i due generi di ambiguità si differenziano sostanzialmente in base alle loro caratteristiche intrinseche: l'ambiguità interna comprende le ambiguità di MSF (genere, numero, persona, ecc.), mentre quella esterna rappresenta l'intera serie di POS assegnate, in fase di *tagging*, a una data forma. Risulta pertanto evidente la possibilità di coesistenza, all'interno delle forme etichettate, di entrambe le ambiguità.

1.1 CARATTERISTICHE SALIENTI DEL LINGUAGGIO DI *SCRIPTING* ADOTTATO**.** Come già accennato in precedenza, AWK è un linguaggio di natura procedurale. Tuttavia le sue caratteristiche interne di funzionamento fanno sì che esso sia uno dei sistemi più semplici, ma nel contempo più efficienti, per la manipolazione di testi. AWK, infatti, dispone di potenti funzioni predefinite quali ad es. la possibilità di realizzare *pattern matching* mediante l'uso di espressioni regolari o la capacità di segmentare un testo intero dividendolo in righe e in campi contenenti i singoli token appartenenti alla riga stessa.

Tuttavia, nel nostro caso, data la natura estremamente *context-sensitive* delle regole di disambiguazione, si è rivelato indispensabile poter operare sul testo con un elevato grado di elasticità. A tal fine, quindi, si è optato per la soppressione della segmentazione automatica del testo in righe successive, in modo da gestire l'intero documento come se fosse costituito da una singola riga intera.

---

[1] Nel prosieguo non saranno commentate le varie “labels” del tagset del CT, per quale basta rimandare al contributo precedente in questo volume, Barbera ¶ 8.

<table>
<tr>
<td>

Il modulo "dis_prep":

```
# Source formatting module
#
{
gsub (/^ /, "")
print $0 "¥"
}
```

</td>
<td>

Il modulo "dis_end":

```
# Format restoring module
#
{
rc = 1
gsub (/\¥ /, "¥")
rec = split ($nf, sp, "¥")
while (rc <= rec)
        {
        print " " sp[rc]
        rc++
        }
}
```

</td>
</tr>
</table>

Tav. 1ab: I moduli "dis_prep" e "dis_end".

Questa soluzione, totalmente priva di svantaggi, ha permesso la creazione di tre puntatori, definiti all'interno del programma dalle variabili "campo", "bw" e "fw". Il primo di essi, "campo", costituisce l'elemento centrale di tutto il sistema di disambiguazione, poiché è preposto alla scansione sequenziale di tutti i token presenti nel testo. Gli altri due puntatori, invece, pur ricoprendo un ruolo importante, possono essere considerati elementi ausiliari in quanto, essendo progettati per esaminare il contenuto del campo immediatamente precedente e immediatamente successivo a quello oggetto di analisi, permettono al linguista di formulare regole contestuali dotate di un notevole grado di precisione. Inoltre l'elevata flessibilità dell'impostazione qui adottata consente, quando necessario, di estendere l'indagine contestuale a una zona di testo anche considerevolmente più ampia rispetto a quella di *default* appena descritta mediante la definizione, all'interno delle regole stesse, di ulteriori puntatori ausiliari. Tuttavia, poiché questa semplice struttura non permette il ripristino della formattazione originale delle righe di testo al termine dell'elaborazione, si è visto necessario affiancare ai 6 moduli costituenti il motore di disambiguazione, due moduli appositamente creati per la gestione dell'aspetto grafico del testo. Il primo di tali moduli, chiamato "dis_prep", cura l'inserimento di un carattere speciale ("¥", scelto arbitrariamente) al termine di ogni linea del testo etichettato originale. Detto carattere funge da marcatore di fine riga, consentendo al secondo modulo di formattazione "dis_end" la fedele ricostruzione del formato grafico originario.

1.2 OTTIMIZZAZIONE DEL SISTEMA. *Last but not least*, allo scopo di restringere l'indagine del disambiguatore unicamente agli elementi testuali considerati linguisticamente rilevanti, si è provveduto al riconoscimento, da parte del sistema, di tutti i codici di markup presenti all'interno delle frasi. Tali codici, del tutto privi di contenuto linguistico, verranno automaticamente saltati dai menzionati puntatori in fase di analisi. Quest'ultimo accorgimento, semplice ma estremamente utile, fa sì che il sistema operi su un testo che può essere considerato a tutti gli effetti 'virtuale' in quanto, ad esclusione dei codici strettamente legati al *tagging* delle varie forme, risulta virtualmente privo di tutte quelle stringhe di caratteri aggiuntive non presenti sul testo cartaceo originale. Può essere ora utile fornire una brevissima analisi delle tecniche di programmazione adottate nello sviluppo del sistema.

Come già accennato in precedenza, l'organizzazione interna dei singoli moduli che formano il disambiguatore è costituita da una serie di regole linguistiche a mutua esclusione. Tuttavia, al fine di ottimizzare al massimo la struttura informatica di tale sistema, si è deciso di sfruttare la caratteristica di AWK che consente la gestione di funzioni definite dall'utente. Una funzione consiste in una parte di codice di programmazione che può essere richiamato, all'interno del programma, da un comando corrispondente al nome della funzione stessa. Al fine di poter sta-

bilire un legame comunicativo tra il corpo del programma e la funzione è necessario che, unitamente al comando di attivazione, vengano forniti una serie di valori denominati “parametri”. La scelta di tali parametri, definita in fase di progettazione, è unicamente vincolata al particolare tipo di elaborazione per cui la funzione è stata predisposta.

L’architettura qui descritta, che, è bene sottolineare, non incide in alcuna misura sui livelli di rendimento computazionale del sistema, offre numerosi vantaggi. Innanzitutto fornisce alle regole linguistiche una maggiore chiarezza espositiva: le regole, essendo meno circondate da linee di programma, potranno essere più facilmente gestibili e modificabili dal personale incaricato anche numerosi anni dopo la conclusione del progetto. Altri vantaggi si riflettono a livello di riduzione delle dimensioni complessive del sistema e di maggiore facilità nella manutenzione della struttura del software.

2 DESCRIZIONE ANALITICA DEGLI ELEMENTI STRUTTURALI COSTITUENTI I VARI MODULI.

Per una migliore comprensione di quanto presentato nei paragrafi precedenti, viene ora fornita una descrizione dettagliata dei blocchi funzionali che si possono incontrare all’interno dei vari moduli. È utile precisare che a parte le funzioni definite dall’utente, tutto ciò che, a livello generale, verrà descritto nel presente paragrafo dovrà necessariamente apparire in ogni modulo. Per quanto riguarda il caso specifico delle funzioni, invece, poiché la scelta della specifica funzione da implementare dipende unicamente dalla complessità computazionale di ciascun modulo, vi saranno moduli in cui potranno coesistere ben quattro funzioni definite dall’utente e moduli in cui una sola funzione risulterà sufficiente per il corretto funzionamento del sistema.

2.1 LINEE DI COMMENTO. Ogni modulo può iniziare con una o più linee di commento in cui vengono indicati il nome del modulo e il tipo di regole ivi ospitate. Tali linee sono immediatamente riconoscibili in AWK in quanto precedute dal simbolo “#”

```
# Motore di disambiguazione - Versione 2.0
#
# Modulo 4:
#       Disambiguazione di:
#
#       - preposizioni, verbi, congiunzioni, ecc.
#
```

Tav. 2: Le linee di commento.

2.2 INIZIO DEL PROGRAMMA. Terminate le righe di commento iniziali, la parte di programma vero e proprio incomincia con una ‘regola’ di programma chiamata “BEGIN”. È necessario puntualizzare che il termine ‘regola’ appena usato non denota una regola linguistica di disambiguazione, bensì una ben precisa procedura inerente al linguaggio di programmazione stesso. AWK richiede che, a parte i comandi “BEGIN”, “END” e le funzioni definite dall’utente, tutte le “regole” che costituiscono un programma siano incluse tra parentesi graffe.

```
BEGIN {
RS = ""
# gestisce l'input come se fosse formato da una riga unica
ORS = " "
# inserisce uno spazio alla fine di ogni 'print'
nf = 1
}
```

Tav. 3: L’inizio del programma

Il comando “BEGIN” viene usato con lo scopo di far eseguire una serie di passi di programma una sola volta all’inizio dell’elaborazione. Nello specifico, in fase di progettazione si è deciso di utilizzare tale comando al fine di definire preventivamente il valore di alcune variabili che verranno usate successivamente all’interno del corpo del programma. In AWK vi sono fondamentalmente due tipi di variabili: le variabili di sistema e le variabili generiche. Le prime, denotate da sigle contenenti solo lettere maiuscole, hanno il potere di modificare impostazioni predefinite o svolgere funzioni particolari; le seconde, invece, definite in genere da lettere minuscole, rappresentano le variabili classiche presenti in ogni linguaggio di programmazione e vengono utilizzate con lo scopo di immagazzinare valori (di tipo numerico o stringa) che possono essere modificati a piacere a seconda delle esigenze. Nel nostro caso specifico, il comando “BEGIN” ci consente di impostare il valore delle variabili di sistema che si occupano della segmentazione del testo in righe. Le variabili in questione, denotate dalle sigle “RS” (*record separator*) e “ORS” (*output record separator*), possono essere programmate al fine di modificare il comportamento standard di AWK così da adattarlo agli scopi dell’utente. Di norma AWK agisce segmentando il testo d’ingresso in righe basandosi sul carattere di fine riga, non visibile, “\n”. In fase di scrittura, invece, il linguaggio inserisce un carattere di fine riga al termine di ogni parte di testo stampata mediante il comando “print”. In accordo con quanto già affermato nel § 1.1, la configurazione appena descritta non risulta adeguata agli scopi del nostro progetto, pertanto si rende necessaria una sostanziale modifica di tale comportamento. Poiché AWK consente di definire, mediante le variabili citate in precedenza, il carattere che l’utente desidera riservare alle funzioni di separatore di riga del testo d’ingresso e separatore di riga in fase di stampa, assegnando alla variabile “RS” un carattere nullo (“”) e ad “ORS” un carattere di spazio (“ ”), si è consentito al disambiguatore di gestire l’intero testo etichettato come composto da una sola riga e di produrre un testo di uscita costituito anch’esso da una sola riga in cui le diverse parti frutto di stampa risultino separate tra loro da uno spazio.

Oltre alle variabili preposte alla gestione della segmentazione delle righe, AWK possiede altre due variabili, “FS” (*field separator*) e “OFS” (*output field separator*). Dette variabili, aventi caratteristiche operative del tutto simili alle precedenti, risultano però responsabili della gestione dei campi. Nel funzionamento di base, i campi contenuti in ogni riga di testo vengono separati tenendo conto della spaziatura. Pertanto, sebbene sia di agevole modifica, questa comportamento viene lasciato del tutto inalterato all’interno dei vari moduli di disambiguazione.

In ultima istanza, nella riga conclusiva del blocco di programma facente capo alla funzione “BEGIN” è stata definita la variabile “nf”, caricata con il valore intero “1”. L’utilizzo di quest’ultima variabile, che descriveremo nel paragrafo successivo, è di importanza fondamentale per il funzionamento stesso del sistema.

2.3 CORPO DEL PROGRAMMA. Le righe iniziali del corpo del programma sono tra le più importanti:

In esse, infatti, si trova la definizione dei tre puntatori cui si fa riferimento nel par. 1.1, l’impostazione delle regole di eliminazione virtuale dei codici testuali non pertinenti (cfr. § 1.2) ed infine il motore di disambiguazione vero e proprio, costituito da regole linguistiche e funzioni definite dall’utente (cfr. § 3 e sg.).

Il funzionamento dell’intero sistema di disambiguazione da noi proposto ruota intorno a un nucleo centrale costituito dalla riga:

```
while (nf <= NF)
```

Nonostante la sua apparente semplicità, tale riga riveste un’importanza fondamentale in quanto è proprio per mezzo di essa che il disambiguatore può procedere al lavoro di scansione all’interno del testo dei vari token ambigui. È doveroso, a questo punto, fornire una descrizione dettagliata di questa linea di codice e del suo funzionamento.

```
{
while (nf <= NF)
    {
#!*!* Inizio regole di disambiguazione *!*!
#
# Creazione di 3 puntatori:
# 'nf' -> punta al campo corrente
# 'bw' -> punta al campo che precede 'nf' di N posizioni
# 'fw' -> punta al campo che segue 'nf' di N posizioni
#
    campo = $nf
    fw = nf
    fw++
    if ($fw ~ /\@/ || $fw ~ /\%/ || $fw ~ /\$/ || $fw ~ /\Ÿ/ || $fw ~ /\#/)
        fw++
    if ($fw ~ /\@/ || $fw ~ /\%/ || $fw ~ /\$/ || $fw ~ /\Ÿ/ || $fw ~ /\#/)
        fw++
    if ($fw ~ /\@/ || $fw ~ /\%/ || $fw ~ /\$/ || $fw ~ /\Ÿ/ || $fw ~ /\#/)
        fw++
# omette le stringhe contenenti:
# '@'
# '%'
# '$'
# 'Ÿ'
# '#'
    bw = nf
    if (nf >=2)
        bw--
if (($bw ~ /\@/ || $bw ~ /\%/ || $bw ~ /\$/ || $bw ~ /\Ÿ/ || $bw ~ /\#/) && bw >
               2)
        bw--
if (($bw ~ /\@/ || $bw ~ /\%/ || $bw ~ /\$/ || $bw ~ /\Ÿ/ || $bw ~ /\#/) && bw >
               2)
        bw--
if (($bw ~ /\@/ || $bw ~ /\%/ || $bw ~ /\$/ || $bw ~ /\Ÿ/ || $bw ~ /\#/) && bw >
               2)
        bw--
# omette le stringhe contenenti:
# '@'
# '%'
# '$'
# 'Ÿ'
# '#'
#
```

Tav. 3: Corpo del programma

Iniziamo con l'analisi del comando "while". Questo comando indica al sistema di ripetere un certo tipo di istruzione, o gruppo di istruzioni, finché la condizione espressa all'interno della parentesi tonda continui a risultare vera. Il ciclo si chiude ed il programma continua il proprio flusso normale solo nel momento in cui la condizione dovesse restituire un risultato negativo, ossia di non verità. Nel nostro caso, quindi, il gruppo di istruzioni incluse nel ciclo "while" verranno ripetute tante volte finché la variabile "nf" non contenga un valore numerico maggiore di "NF". È evidente, quindi, come la procedura di aggiornamento di "nf" ricopra un ruolo delicato: se non ben realizzata, può presentarsi il rischio di un ingresso in *loop* dell'esecuzione del programma (caratterizzato dalla ripetizione all'infinito dello stesso comando) o, in alternativa, possono risultare alcune perdite di dati nel testo di uscita. Per ovviare a tali rischi, pertanto, il valore contenuto in "nf" viene aggiornato dal programma immediatamente dopo l'analisi di ciascun elemento testuale. Se riguardo a "nf" non vi è molto da aggiungere a quanto già detto fino-

ra, la variabile “NF” richiede invece un commento più articolato. Come già accennato in precedenza, il linguaggio di programmazione da noi adottato utilizza al suo interno una serie di variabili di sistema dedicate allo svolgimento di compiti ben precisi. La variabile “NF” (*Number of Field*) è anch’essa una variabile di sistema che però, a differenza di quelle già incontrate, fornisce il conteggio della quantità di campi presenti all’interno del testo. Poiché nel nostro sistema i campi vengono divisi tenendo conto del carattere di spazio, “NF” fornirà il valore corrispondente alla quantità di token presenti nel testo da analizzare.

Date tali premesse, diventa più agevole comprendere la riga di programma presentata: finché la variabile generica “nf”, inizialmente caricata con il valore numerico ‘1’, conterrà un valore minore od uguale al numero totale dei campi contenuto in “NF”, il sistema procederà all’esecuzione delle varie regole di disambiguazione presenti all’interno del ciclo “while”. La scansione del testo si interromperà, invece, solo nel momento in cui “nf” conterrà un valore maggiore di “NF”, segno che anche l’analisi dell’ultimo token ha trovato compimento.

In AWK, come in altri linguaggi quali il C, C++, Java, Perl, ecc. è necessario l’uso delle parentesi graffe per includere quelle parti di programma che risultano gerarchicamente dipendenti da altre. Pertanto il corpo delle regole di disambiguazione, dipendendo direttamente dal precedente comando “while”, dovrà essere preceduto da una parentesi graffa aperta.

Proseguendo con la descrizione analitica del programma, ci accingiamo ora ad esaminare nel dettaglio la definizione dei puntatori “bw”, “fw” e “campo”. I tre puntatori qui elencati si trovano all’interno del gruppo di istruzioni che, gerarchicamente dominate dal “while” di cui sopra, costituiscono il sistema di disambiguazione vero e proprio. Il puntatore “campo”, infatti, è una variabile definita dalla riga:

```
campo = $nf
```

Tale linea di programma fa sì che all’interno di “campo” venga caricata la stringa di caratteri appartenente al campo indicato dal valore di “nf”. Il simbolo “$” che precede “nf” indica appunto che “campo” conterrà un valore di tipo stringa e non di tipo numerico.

Gli altri due puntatori, invece, partendo sempre dal valore di “nf”, consentono di leggere il contenuto del testo presente nei campi immediatamente precedenti ed immediatamente successivi a “campo”. Tuttavia in questo contesto si inserisce anche il sistema di controllo automatico dei codici di markup, elementi testuali totalmente privi di rilevanza in seno al processo di disambiguazione. Tale sistema automatico prevede l’incremento del valore contenuto nella variabile “nf” ed “fw” ed il decremento di “bw” ogniqualvolta il sistema incontri un campo in cui siano presenti i simboli: “@”, “%”, “$”, “*f*” e “#”. Come già accennato in precedenza, questo accorgimento consente di elaborare regole di disambiguazione che agiscono su materiale puramente testuale, senza dover tenere conto di tutti gli elementi di natura extralinguistica presenti nel testo etichettato. Per maggiore completezza descrittiva è bene precisare che solo “nf”, in quanto variabile centrale, subirà un incremento pari ad uno. Le altre due variabili “fw” e “bw”, invece, in virtù della loro funzione ausiliaria, potranno subire variazioni differenti, in stretta relazione con il numero di codici di markup che è necessario saltare prima di incontrare un elemento di testo valido. Poiché il testo può presentare i suddetti codici in posizione consecutiva fino a un massimo di quattro, si è predisposto un sistema di controllo per evitare che “bw” possa assumere valori negativi, rischio presente soprattutto nei momenti iniziali dell’elaborazione.

3 REGOLE DI DISAMBIGUAZIONE. Non potendo, per ovvie ragioni di spazio, presentare un’analisi completa di ciascuna delle regole linguistiche implementate nel sistema, ci limiteremo ad un excursus parziale prendendo in esame alcune delle regole più significative presenti nei vari moduli. Prima di addentrarci nell’argomento, però, è opportuno precisare che, poiché all’interno di un modulo le varie regole sono organizzate in un sistema sequenziale a mutua esclusione, queste dovranno essere disposte tenendo conto del loro livello di generalizzazione.

Una regola che agisce prendendo in esame i valori di HDF ed MSF sarà dotata di una capacità di disambiguazione nettamente più ampia e generale rispetto ad una regola che basa la sua capacità di azione unicamente sull'analisi del lemma o della forma di un dato token. Date queste premesse, risulta chiaro che la presenza in uno stesso modulo di due regole differenti che trattano una problematica comune (es. le forme straniere), richiederà uno studio accurato sulla loro dislocazione all'interno del modulo stesso, al fine di evitare che l'entrata in funzione di una determinata regola *ad hoc* (ossia *context sensitive*) venga impedita dalla compresenza di una regola generale di tipo *context free*.

Una norma che consente di ottenere una certa sicurezza organizzativa consiste nel disporre le regole dotate di maggiore generalizzazione in una posizione più avanzata rispetto a quelle legate al contesto specifico, che saranno pertanto le prime ad entrare in azione. Questo aspetto, che incide in primo luogo sull'organizzazione interna, si riflette anche a livello esterno sulla disposizione sequenziale dei moduli: quelli caratterizzati dal possedere regole generali, infatti, entreranno in funzione solo in un momento successivo rispetto ai moduli costituiti da regole sensibili al contesto. Tuttavia, è bene precisare che la scelta del tipo di regole da inserire all'interno dei vari moduli è anche strettamente legato alla capacità di analisi che si intende attribuire ai moduli stessi. Se si prende in esame, in qualità di esempio, il sistema di regole adottato per il trattamento degli articoli determinativi transcategorizzanti con pronomi, è possibile notare che, a differenza di quanto detto poc'anzi, le regole di portata generale sono presenti in un modulo antecedente a quello che contiene le regole che agiscono ad un livello più specifico. Questo tipo di scelta, apparentemente in contrasto con i principi base di ortodossia organizzativa, trova la sua giustificazione nel fatto che i risultati di questa specifica azione di disambiguazione, che richiede un sistema di analisi piuttosto complesso ed articolato, possano essere immediatamente utilizzati da altre regole presenti nei moduli immediatamente successivi. Mediante tale disposizione, infatti, la disambiguazione avviene in due moduli ed in due momenti ben precisi e distinti: il primo gruppo di regole, infatti, agisce nel terzo modulo di programma e si comporta come un filtro a maglia larga, occupandosi quasi unicamente di discriminare gli articoli determinativi dalle corrispettive forme pronominali. Il secondo gruppo, invece, che agisce nel quarto modulo, si occupa più nello specifico di assegnare loro i corretti valori di lemma. Poiché numerose regole richiedono la disambiguazione dell'articolo o del pronome per poter portare a termine il proprio compito, appare evidente come l'importanza di una discriminazione, seppur grossolana, della POS sia nettamente prioritaria rispetto al compito di assegnazione del lemma corretto; da qui la scelta, quasi obbligata, di una organizzazione delle regole in una maniera che può apparire, a prima vista, alquanto irrazionale. In conclusione, ritornando al discorso riguardante l'importante aspetto dell'organizzazione interna del sistema di regole, possiamo comunque ragionevolmente affermare che è sempre consigliabile optare, ogniqualvolta si presenti la possibilità, verso l'accorpamento, nei diversi moduli, delle regole con caratteristiche comuni, in modo da evitare il più possibile la promiscuità tra tipi di regole caratterizzate da capacità di analisi differente.

3.1 ESEMPIO DI REGOLA TRATTA DA "MODULO 1". Formato unicamente da regole di tipo *context sentitive*, il modulo 1 è interamente dedicato al trattamento dei casi di ambiguità verbale interna e/o esterna non risolvibili mediante regole generali.

In tavola 4 se ne fornisce un esempio (in corpo ridotto per economia di spazio), che sarà poi partitamente analizzato.

A	# Regola per la disambiguazione interna
	# ed esterna della forma 'ave'
	else
	if (campo ~ /^ave_/ && campo ~ /\);\(/)
	{

B	if (campo ~ /¥$/)
	end = "¥"
	else
	end = ""

C	nf++

D	if ($fw ~ /^÷lle_/)
	{
	assegna(campo, "221", end)
	}
	else
	if ($fw ~ /^mari[ae]_/)
	{
	assegna(campo, "68", end)
	}

E	else
	if ($fw ~ /^÷gli_/)
	{
	sub (/;3/, "", campo)
	sub (/6;/, "", campo)
	assegna(campo,"211",end)
	}

F	else
	{
	sub (/2;/, "", campo)
	sub (/;7/, "", campo)
	assegna(campo,"211",end)
	}
	}

Tav. 4a-f: Una regola di disambiguazione del modulo 1

A è l'elemento di controllo che si occupa di verificare la possibilità dell'entrata in funzione della regola mediante l'esecuzione di un confronto (*pattern matching*) tra il valore di stringa contenuto nella variabile "campo" e lo specifico token che la regola intende trattare. La richiesta di un'operazione di confronto tra modelli di stringhe viene inoltrata al linguaggio AWK mediante l'uso del simbolo speciale "~".

B introduce ulteriori elementi di controllo finalizzati alla corretta gestione del marcatore di fine riga (cfr. § 1.1).

C è la linea riservata all'incremento della variabile "nf" che scansiona il testo (cfr. § 1.1).

D è la porzione di regola che rappresenta l'aspetto *context sensitive* del disambiguatore: utilizzando il confronto tra la stringa contenuta nel campo successivo e quella necessaria per poter assegnare un determinato valore di POS, la regola comanda al sistema di eseguire l'operazione di eliminazione dell'ambiguità esterna. Tale ordine viene impartito ricorrendo alla funzione "assegna", alla quale devono essere comunicati i parametri necessari per lo svolgimento del lavoro di disambiguazione vero e proprio (cfr. § 4).

E, poi, è la parte di regola che, oltre alla funzione descritta nel punto precedente, comprende anche la gestione dell'ambiguità interna. Questa viene eliminata ricorrendo al comando "sub" (*substitution*), funzione che consente di modificare un determinato valore alfanumerico all'interno di una variabile stringa. In dettaglio, la disambiguazione interna viene ottenuta sostituendo all'interno di "campo" il valore di MSF non desiderato con un carattere nullo.

F, infine, è il finale della regola, costituito in questo specifico caso unicamente da comandi per la disambiguazione interna, indica al sistema il comportamento a cui attenersi nel caso in cui i precedenti controlli sui campi circostanti dovessero dare esito negativo. Il finale di regola qui descritto è importante poiché consente di evitare la formulazione di regole specifiche necessarie a coprire tutta l'ampia casistica di variazione del contesto, pertanto è presente in quasi tutte le regole appartenenti ai vari moduli.

4 FUNZIONI DEFINITE DALL'UTENTE. Riguardo a questo argomento il manuale di AWK afferma che «Definitions of functions can appear anywhere between the rules of an 'awk' program», ossia le funzioni definite dall'utente possono trovarsi ovunque tra le regole di programma. Questa caratteristica, che volendo consente al programmatore di inserire le funzioni anche

al fondo dell'intero listato di codice, è data dal fatto che questo linguaggio di programmazione esamina preventivamente l'intero programma prima di procedere all'esecuzione. Pertanto noi tratteremo il presente argomento come una sorta di entità autonoma e separata rispetto al corpo del programma vero e proprio.

In AWK una funzione si dichiara usando il comando "function" seguito dal nome della funzione stessa. Esso è a sua volta seguito da una parentesi tonda contenente i parametri (cfr. § 3.1) e le variabili che operano all'interno della funzione. Le varie funzioni del nostro programma sono caratterizzate dall'avere un numero di parametri costante, ma un numero di variabili differente. Occorre infine precisare che il carattere di spazio che separa i due blocchi di elementi all'interno della parentesi tonda è totalmente privo di qualsiasi utilità computazionale: il suo utilizzo viene consigliato unicamente per favorire la leggibilità del programma.

Le funzioni definite dall'utente costituiscono, nel nostro sistema, il motore vero e proprio del sistema di disambiguazione. È al loro interno, infatti, che avviene il processo di selezione ed assegnazione della categoria grammaticale corretta e l'eliminazione di tutte le altre transcategorizzazioni superflue.

Per una migliore comprensione del processo di disambiguazione, riportiamo in Tav. 5 le linee di programma riferite alla funzione "assegna", seguite dalla relativa descrizione analitica.

```
A | function assegna(campo, pos, end, cpn, cp, csp, sp, spl, cl)
  |     {
B |     cpn = 1
C |     cp = split (campo, sp, /\);\(/)
D |     csp = split (sp[1], spl, /\(/)
E |     pos = pos ","
F | while (cpn <= cp)
  |         {
G |         if (cpn > cp)
  |             break
H |         if (sp[cpn] ~ pos)
  |             {
I |         if (sp[cpn] ~ /\)$/ || sp[cpn] ~ /\)¥$/)
  |                 {
  |                 cl = sp[cpn]
  |                 sub (/\)/, "", cl)
  |                 print spl[1] cl
  |                 }
  |             else
  |             if (cpn == 1)
  |                 print spl[1] spl[2] end
  |             else
  |                 print spl[1] sp[cpn] end
  |             }
J |             cpn++
  |         }
  |     }
```

Tav. 5a-j: La funzione "assegna"

A contiene la dichiarazione della funzione, dei parametri e delle variabili adottate e B la dichiarazione della variabile "cpn" ed assegnazione del valore numerico "1".

C fa uso della funzione predefinita “split” al fine di separare le varie transcategorizzazioni inserendo i diversi valori di POS in una tabella (*array*).

In D, poi, si utilizza di nuovo “split” per separare il token dal gruppo di transcategorizzazioni.

In E si inserisce un segno di virgola al termine della stringa di caratteri numerici convogliata dal parametro “pos”.

In F, per mezzo del comando “while” e l’uso della variabile “cpn”, si istituisce un ciclo iterativo per scansionare le varie POS presenti nell’*array* precedentemente costituito.

G, quindi, verifica il punto di scansione per l’interruzione al momento opportuno del ciclo iterativo; e H seleziona la categoria corretta mediante il confronto tra il contenuto del parametro “pos” e le POS transcategorizzanti oggetto di scansione.

I, in caso di esito positivo del confronto, ricostruisce e stampa su file la nuova linea di testo etichettata. Il simbolo “¥” viene utilizzato al fine di permettere, al termine dell’elaborazione del modulo finale, il ripristino della formattazione del testo del file originale.

J, infine, in caso di esito negativo, incrementa la variabile “cpn” e continua il ciclo iterativo di scansione.

BIBLIOGRAFIA.

ARMSTRONG
1994 *Using Large Corpora*, edited by Susan Armstrongs, Cambridge (Mass.) - London (En.), The MIT Pres, 1994 “A Bradford Book”, “ACL-MIT Press Series in Computational Linguistics” [= “Computational Linguistics” XIX (1993)[1-2]].

ATWELL - SOUTER 1993 → SOUTER - ATWELL 1993

BARBERA
¶ 1 Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in questo volume, pp. 3-20.
¶ 8 Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in questo volume, pp. 135-168.

BRENNAN
2000 Michael Brennan, *GAWK: Effective AWK Programming: A User’s Guide for GNU Awk*, 2nd edition, Free Software Foundation Inc., 2000, disponibile online alla pagina: `http://www.gnu.org/software/gawk/manual/gawk.html`.

GUTHRIE
1993 Louise Guthrie, *A Note on Lexical Disambiguation*, in SOUTER - ATWELL 1993, pp. 11-24.

HINDLE - ROOTH
1994 Donald Hindle - Mats Rooth, *Structural Ambiguity and Lexical Relations*, in ARMSTRONG 1994, pp. 103-120.

MASON
2000 Mason, Oliver, *Programming for Corpus Linguistics - How to Do Text Analysis with Java*, Edinburgh, Edinburgh University Press, 2000 “Edinburgh Textbooks in Empirical Linguistics”.

MATTHEWS
1998 Matthews, Clive, *An Introduction to Natural Language Processing through Prolog*, New York, Longman, 1998 “Learning about language”.

MITKOV
2003 *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov, Oxford, Oxford University Press, 2003.

SOUTER - ATWELL
1993 *Corpus-based Computational Linguistic*s, edited by Clive Souter and Eric Atwell, Amsterdam - Atalanta, Rodopi, 1993 "Language and Computers: Studies in Practical Linguistics" 9.

STEVENSON - WILKS
2003 Mark Stevenson - Yorick Wilks, *Word-Sense Disambiguation*, in MITKOV 2003, pp. 249-265.

VOUTILAINEN
2003 Atro Voutilainen, *Part-of-Speech Tagging*, in MITKOV 2003, pp. 218-232 [soprattutto § 11.6 *Handwritten Disambiguation Grammars*, pp. 227-230].

WEISCHEDEL et alii
1994 Ralph Weischedel - Marie Meteer - Richard Schwartz - Lance Ramshaw - Jeff Palmucci, *Coping with Ambiguity and Unknown Words through Probabilistic Models*, in ARMSTRONG 1994, pp. 319-342.

CORPORA DI RIFERIMENTO.

Corpus Taurinense → CT.

CT `http://www.bmanuel.org/projects/ct-HOME.html`.

ANGELA FERRARI - MAGDA MANDELLI
(Universität Basel)

10. Note sull'impiego dei connettivi nei notiziari accademici del corpus *Athenaeum*. *Aspetti quantitativi e qualitativi.*

0. INTRODUZIONE. Nel corpus *Athenaeum* riunito dall'équipe di ricercatori torinesi guidata da Carla Marello[1] compare un vasto insieme di testi tratti dalla rivista *L'Ateneo*, il notiziario accademico dell'Università di Torino. Malgrado la comune matrice, si tratta di scritture tipologicamente eterogenee, che si differenziano dal punto di vista della loro funzione retorico-illocutiva – ve ne sono di strettamente espositive, di esplicative, di argomentative –, della natura denotativa dei loro argomenti e del rapporto che esse intrattengono con l'orale: accanto a testi scritti in senso stretto, vi sono infatti anche testi che sono stati scritti per essere detti nel corso di celebrazioni o di riunioni ufficiali di vario genere.

Sulla base di un campione significativo di articoli, in Ferrari 2005b è stato proposto un paradigma di osservazioni relative ai testi "scritti-scritti" di *L'Ateneo* con funzione espositiva (od «infomativa», come dice l'etichettatura che qualifica i vari testi nel *corpus*) od espositivo-esplicativa. L'obiettivo consisteva, più precisamente, nell'identificare le peculiarità logico-testuali (cfr. *infra*) della tipologia prescelta, sintomatiche della sua specificità di "notizia accademica", specificità che combina la funzione retorico-illocutiva di esposizione con il carattere "accademico" dei temi trattati e dei partecipanti all'atto comunicativo. In questa sede, ripercorreremo le conclusioni proposte in Ferrari 2005b osservandole – al fine di confermarle, arricchirle, modularle – alla luce di un'analisi, sia quantitativa sia qualitativa, più precisa e sistematica dei connettivi[2]. Il campione qui considerato è costituito da un insieme di testi di *L'Ateneo* per un totale di circa 60.000 parole che interseca il campione affrontato in Ferrari 2005b.[3]

L'analisi, oltre a tratteggiare alcune caratteristiche della tipologia testuale affrontata, permetterà anche di segnalare brevemente l'insieme di contrassegni linguistici e testuali necessari, in generale, a fissare le peculiarità di un tipo di testo rispetto all'altro (cfr. Ferrari 2005a); e per approfondire quella sottoclasse di indizi tipologici che disegnano la "trama logica" dei testi.

---

[1] L'allestimento del *corpus*, etichettato dal punto di vista grammaticale, si colloca nel quadro del progetto FIRB.
[2] I dati sistematici relativi ai connettivi sono stati elaborati da Magda Mandelli e presentati sotto la forma di poster nell'incontro di studio *Corpora e linguistica in rete* (Torino, 30 settembre 2005).
[3] Qui di seguito i titoli degli articoli analizzati: *Ricerche etnologiche e accordi di cooperazione dall'Africa Equatoriale all'Africa Occidentale; L'Africa e il centro per lo studio delle Letterature e delle culture delle aree emergenti; L'Università di Torino e l'Africa letteraria di espressione francese; Attività della missione archeologica italiana a Abuqir (Egitto); Gli scavi archeologici di Abuqir*; *Sostenibilità ambientale, sostenibilità umana. Alcune esperienze africane di ricerca-azione; La facoltà di medicina veterinaria e l'Africa; L'attività del CNR nel settore amianto tra passato e futuro; Epidemiologia delle malattie da amianto in Italia*; *Ai confini della realtà: storie della teoria della stringa; La rifondazione dello Studio torinese: Vittorio Amedeo II e l'Università; Il settecentesco Palazzo degli Studi; Il cantiere di restauro; Il restauro del Palazzo dell'Università di Torino; L'analisi dei trattamenti murali negli stucchi e decorazioni murarie; La ricerca storica quale strumento finalizzato al restuaro; La ricerca nei laboratori scientifici «A. Mosso»; Proposta di recupero dell'istituto scientifico «A. Mosso» al Col D'Olen (Monte Rosa). Un museo della scoperta scientifica del Monte Rosa; Animali africani nel museo di zoologia dell'Università di Torino; La mostra «L'Africa in Piemonte tra '800 e '900»; Recenti sviluppi della «questione amianti in Italia».*

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 183-198.

1. L'ARCHITETTURA "LOGICA" DEL TESTO COME CONTRASSEGNO TIPOLOGICO. Una tipologia testuale che sia interessante per le scienze linguistiche deve essere correlata a particolari e distintive proprietà legate all'espressione verbale (ad es. Sabatini 1998; Mortara Garavelli 2001). Per quanto concerne la comunicazione scritta, tali proprietà, che diventano dunque i "contrassegni" di ogni singolo tipo di testo, riguardano il lessico; la costruzione morfosintattica ed interpuntiva della clausola intesa come unità massimale della lingua; e le strategie di strutturazione testuale (per una precisazione e illustrazione di questo sistema di contrassegni tipologici, cfr. Ferrari 2005a, pp. 16-38).

1.1 LA GESTIONE DEI CONTENUTI. Da quest'ultimo punto di vista, sono particolarmente indicative la gestione dei contenuti semantico-pragmatici del testo come significati espliciti od impliciti, cioè ricostruiti inferenzialmente sulla base di dati extralinguistici; e l'organizzazione dei contenuti espliciti entro l'architettura che coglie l'essenza della testualità.

Si tratta di un'architettura complessa, che organizza le unità gerarchiche costitutive del testo (unità informative, enunciati, gruppi di enunciati, capoversi ecc.) entro un insieme definito di dimensioni semantico-pragmatiche: la dimensione che ruota attorno al concetto di *topic*, caratterizzato in termini di *aboutness* (nel senso di Lambrecht 1994); quella incentrata sulle relazioni logiche (motivazione, esemplificazione, riformulazione ecc.); la dimensione che coglie cambiamenti composizionali quali ad esempio il passaggio dalla narrazione alla descrizione, da questa all'argomentazione, e così via); la dimensione che misura l'intrecciarsi nel testo dei diversi punti di vista; ecc.

1.2 LA STRUTTURAZIONE LOGICA. La strutturazione logica del testo – su cui si concentreranno le pagine seguenti – è dunque una componente del paradigma di contrassegni pertinenti per una caratterizzazione testuale significativa nell'àmbito della *Textsortenlinguistik*. Essa si definisce più precisamente attraverso la fissazione delle variabili associate ai seguenti parametri (cfr. Ferrari 2005b, pp. 246-270):

(j) *Primo parametro*: la natura concettuale delle relazioni logiche (motivazione, consecuzione, riformulazione ecc.)
(ij) *Secondo parametro*: il carattere esplicito od implicito delle unità connesse e della loro articolazione logica
(iij) *Terzo parametro*: la distanza delle unità connesse
(iiij) *Quarto parametro*: i livelli dell'articolazione logica (relazioni logiche tra unità informative, enunciati, capoversi ecc.)
(v) *Quinto parametro*: l'espressione linguistica delle relazioni logiche (relazioni logiche segnalate o non segnalate linguisticamente, natura morfosintattica e semantica dei segnali prescelti ecc.)

L'interesse tipologico dell'organizzazione logica del testo può riguardare altrettanto bene: le modalità secondo le quali si concretizza ognuno di questi parametri, i modi dell'interazione dell'insieme di questi parametri, le strategie in funzione delle quali il complesso dell'organizzazione logica dialoga con gli altri tipi di strutturazione semantico-pragmatica (topicale, composizionale, enunciativo-polifonica ecc.), lasciando loro spazio od al contrario sottraendoglielo.

2. I CONNETTIVI. Fondamentalmente in linea con Pasch - Brauße - Breindl 2003, § A (a cui si deve la più importante trattazione "formale" e semantica dei connettivi), etichettiamo come connettivi quelle espressioni linguistiche che (j) non sono soggette a flessione morfologica e (ij) indicano una connessione logica – motivazione, consecuzione, concessione, esemplificazione ecc. – tra due (o più) entità semantiche associate minimalmente ad uno stato di cose, associa-

te cioè a unità "ontologicamente" superiori quali gli stati di cose valutati epistemicamente od illocutivamente, ma non ad entità di primo grado.

2.1 CONSISTENZA GRAMMATICALE. Questa definizione, che coniuga un criterio morfologico con un criterio semantico-funzionale, ha un'ampia estensione linguistica.

Le condizioni (j) e (iji) sono infatti soddisfatte dalle congiunzioni coordinanti che articolano nominalizzazioni sintagmatiche, clausole od unità sintattiche superiori; da tutte le congiunzioni subordinanti riconosciute dalla tradizione grammaticale; da tutte le espressioni avverbiali o congiuntive (*tuttavia*, *dunque*, *nondimeno*, *per esempio*, *infatti*, *tutto sommato*, *vale a dire, in particolare* ecc.) con valore logico-relazionale; da quelle preposizioni e le locuzioni preposizionali (*a causa di, malgrado* SN, *eccetto* SN ecc.) che reggono sintagmi nominali la cui testa è un nome argomentale,

[1] // la riunione non è potuta cominciare **a causa d**el suo ritardo//[4],

così come da forme che introducono clausole non temporalizzate (*al fine di*, *per* ecc.).

Non è invece un connettivo l'espressione *dietro* in un esempio come [2], in quanto essa qualifica un'entità di primo grado, né direttamente né indirettamente eventiva:

[2] // Giovannino si è nascosto **dietro** la casa//.

Il testo seguente, che riprendiamo da Ferrari 2005b, p. 267, illustra la varietà linguistica con cui si manifesta la categoria morfosemantica dei connettivi:

[3] In quest'ultimo caso, **vale a dire** nella situazione in cui si trova l'italiano dal Cinquecento ad oggi, possiamo ulteriormente riconoscere dei a) *periodi a normazione rigida*, e dei b) *periodi a normazione debole*. Per fare qualche esempio, possiamo dire che tutto il XVI secolo è stato un periodo a normazione rigida, almeno a partire dal 1525, anno di pubblicazione delle *Prose della volgar lingua* del Bembo, l'atto di fondazione della lingua letteraria comune; **così come** sono stati gli ultimi trenta anni del XIX secolo, **dopo** l'accettazione su scala nazionale della riforma linguistica manzoniana **e** il proliferare di grammatiche ispirate ad essa. Il Novecento, **al contrario**, è un secolo a normazione debole: lo dimostra la scarsa produzione di grammatiche normative, **in rapporto con** la relativa stabilità dello standard linguistico a livello di strutture fonetiche e morfologiche, **mentre** qualche innovazione di rilievo si ha nella sintassi e soprattutto nell'incremento del lessico neologico. Tesi 2001, pp. 8-9.

2.2 STRUTTURAZIONE DEL TESTO. Si noti che, oltre che attraverso i connettivi, i concetti relazionali che strutturano il testo dal punto di vista logico possono essere veicolati da sintagmi nominali (***la causa** è che*...) ed espressioni verbali, come mostra la sequenza [4] tratta da [3],

[4] Il Novecento, al contrario, è un secolo a normazione debole: lo **dimostra** la scarsa produzione di grammatiche normative, in rapporto con la relativa stabilità dello standard linguistico a livello di strutture fonetiche e morfologiche [...] Tesi 2001, p. 9,

da clausole vere e proprie, od essere in forma di (pseudo-)subordinata (sempre a partire da [3]),

[5] **Per fare qualche esempio**, possiamo dire che tutto il XVI secolo è stato un periodo a normazione rigida, almeno a partire dal 1525, anno di pubblicazione delle *Prose della volgar lingua* del Bembo, l'atto di fondazione della lingua letteraria comune [...] Tesi 2001, p. 8,

od in forma di enunciato autonomo, come nei casi delle espressioni *Ecco qualche esempio* o *Vediamo ora un esempio*.

---

[4] Salvo diversamente avvisato, gli esempi in Courier sono tratti dall'Athenaeum Corpus; in Times, invece, sono gli *exempla ficta* e quelli tratti da altre fonti (segnalate).

2.3 SEMANTICA. Dal punto di vista della loro semantica, non tutti i connettivi sono caratterizzati dalla stessa ricchezza e dalla stessa univocità semantiche. Accanto ad espressioni piene e semanticamente rigide come *vale a dire*, *al contrario, benché* od *a condizione che*, ci sono espressioni concettualmente caratterizzate ma più flessibili come *perché* o *quando*; espressioni ambigue (cioè provviste di almeno due valori in rapporto di esclusione) come *ovvero*, che può avere valore disgiuntivo o riformulativo; ed espressioni semanticamente povere quali – in un crescendo di sottospecificazione – la "preposizione" *per*, la congiunzione *se*, la congiunzione *e* (Ferrari 2005b, p. 267).

La varietà di forme con cui possono essere segnalate le relazioni ha un'importante incidenza sui modi in cui si definisce l'architettura logica del testo. Innanzitutto nella misura in cui ogni variazione lessicale porta con sé specificità semantico-logiche che possono rivelarsi cruciali per una caratterizzazione tipologica dei testi: basti pensare ai diversi modi in cui si colora una motivazione quando è espressa da *perché*, *dato che*, *siccome*, *infatti*, *tanto più che*, *la ragione è che* ecc.; od ancora, restando questa volta nell'àmbito della stessa categoria morfosintattica, ai differenziali semantici di locuzioni subordinanti condizionali quali *se*, *a patto che*, *a condizione che* o *sempre che* ecc.: cfr. Visconti 2000). In secondo luogo per il diverso tipo di "testualizzazione" – pensiamo in particolare alla portata ed al rilievo attribuiti alla connessione logica – che tale varietà linguistica implica (cfr. Ferrari 1999, Ferrari 2006a). Così per esempio, segnalare proletticamente una motivazione con un intero enunciato – *vs* scegliere una congiunzione subordinante od un elemento avverbiale – vuol dire anzitutto dare al movimento una particolare importanza nella gestione della "logica" del testo; in secondo luogo suggerire una certa complessità e ampiezza del movimento testuale a venire; e in terzo luogo – qualora il nucleo della clausola fosse arricchito con elementi circostanziali e aggettivi con funzione aggiuntiva: *la vera causa di questo fatto, generalmente ignorata dai più* ecc. – modulare natura e punto di vista della motivazione.

Data una stessa forma linguistica, incide inoltre sui modi della testualizzazione del movimento logico anche la manifestazione sintattico-interpuntiva del connettivo e dei suoi connessi. Così, come si mostra ampiamente in Ferrari 2004a, Mandelli 2004 e Ferrari - Mandelli *i.p.*, la stessa congiunzione subordinante o coordinante può creare rilievi fortemente differenziati in funzione della sua distribuzione sintattica e della punteggiatura che la accompagna: l'operando interno dei connettivi (*perché* ed *e* in particolare) sarà per esempio sullo sfondo informativo dell'enunciato se compare in posizione sintatticamente inserita; nel caso della subordinazione, esso sarà invece in primo piano, potendo diventare addirittura il Fuoco comunicativo dell'enunciato, se linearizzato in posizione conclusiva. E lo stesso tipo di analisi si applica, *mutatis mutandis*, alle locuzioni avverbiali non subordinanti (Ferrari 2005, Mandelli *i.p.*, Ferrari *i.p.*): se esse saturano la posizione incipitaria dell'enunciato hanno un rilievo testuale maggiore rispetto a quando si manifestano in inserzione sintattica tra due virgole, differenza che può anche incidere sull'interpretazione logico-semantica del connettivo.

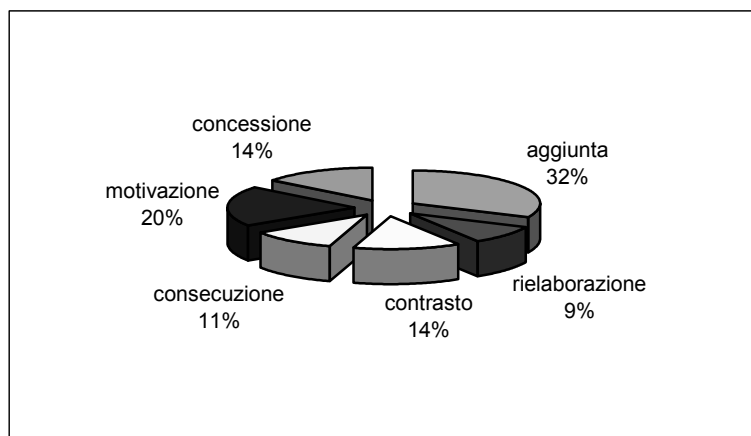
2.4 TIPOLOGIA DEI TESTI. Ai fini di una caratterizzazione tipologica dei testi, l'analisi dei connettivi svolge un ruolo particolarmente importante. Lo svolge di per sé, nella misura in cui per esempio un'ampia variazione nella scelta delle loro forme sintattico-lessicali – soprattutto se interna ad una stessa funzione logico-semantica – è il segno della scelta di un registro elevato e controllato, tipica manifestazione di una varietà di lingua (detta) standard-letteraria. Ma lo svolge anche in quanto essa ci dà informazioni su aspetti più "nascosti", meno immediatamente visibili dell'organizzazione del testo. Ciò vale in particolare per quella dimensione testuale che abbiamo chiamato "logica". Se è vero che essa si può manifestare, e definire attraverso i parametri visti sopra, anche in assenza di connettivi, è altrettanto vero che i connettivi – per la loro natura morfosintattica intrinseca, per la loro distribuzione sintattico-interpuntiva, per i valori

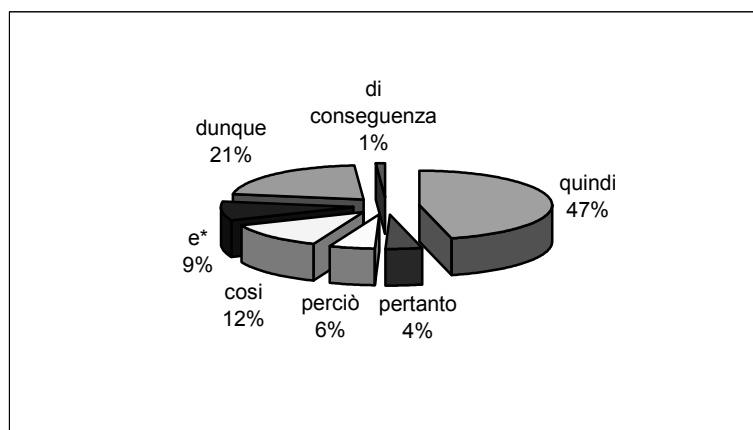
logici che attivano – sono sintomi trasparenti e preziosi delle modalità semantico-pragmatiche in cui essa si realizza.

Sullo sfondo di questo assunto, noi, come già detto nell'introduzione, ripercorreremo gli aspetti principali dell'architettura logica dei notiziari accademici di *L'Ateneo* tratteggiata in Ferrari 2005b osservandoli alla luce dei connettivi, di una loro analisi quantitativa e qualitativa più precisa e sistematica.

3. CONNETTIVI E NATURA CONCETTUALE DELLE RELAZIONI LOGICHE. Esaminiamo ora la natura "concettuale"' delle relazioni logiche caratteristiche dei notiziari accademici.

3.1 LE RELAZIONE LOGICHE. Le relazione logiche su cui si fonda più caratteristicamente la strutturazione logica dei contenuti semantico-pragmatici di un testo scritto possono essere raggruppate nelle seguenti macroclassi: la relazione di aggiunta; di concessione-limitazione; di motivazione (inclusa la finalità); di consecuzione; di contrasto e di rielaborazione semantica e "formale" (riformulazione parafrastica, illustrazione, esemplificazione, particolarizzazione, generalizzazione). Partendo da questa classificazione, in Ferrari 2005b si osservava, ragionando in negativo, che vi erano due tipi di relazione poco rappresentati: il tipo consecutivo e il tipo rielaborativo; dato, quest'ultimo, tanto più degno di nota in quanto la rielaborazione testuale riunisce relazioni caratterizzate da una certa varietà semantica. Questi due dati non sono a ben guardare sorprendenti: la bassa presenza dei due macrotipi di relazione si spiega infatti alla luce della natura pragmatica dei testi considerati. La debole frequentazione della consecuzione va ricondotta al carattere globalmente espositivo-esplicativo dei testi: questa tipologia retorico-illocutiva, in particolare quando non è accompagnata da obiettivi didattici, predilige spiegare dati ed ipotesi riccorrendo alla movenza logica della motivazione (*perché*, *infatti* ecc.); la relativa scarsa presenza di relazioni di rielaborazione risiede, invece, nella natura tendenzialmente essenziale e compatta dell'esposizione e nell'omogeneità socio-professionale di destinatori e destinatari.

L'osservazione *e negativo* proposta in Ferrari 2005b risulta confermata dalla distribuzione dei connettivi nelle diverse classi logico-semantiche osservata nel nostro campione:

Tav. 1: La distribuzione "logico-semantica" dei connettivi.

Se – come è necessario fare – si va al di là dei dati grezzi di natura quantitativa, il risultato si fa ancora più marcato. Per quanto riguarda la consecuzione, osservando più da vicino la semantica dei connettivi, ragionando cioè sulle percentuali seguenti

Tav. 2: I connettivi di consecuzione.

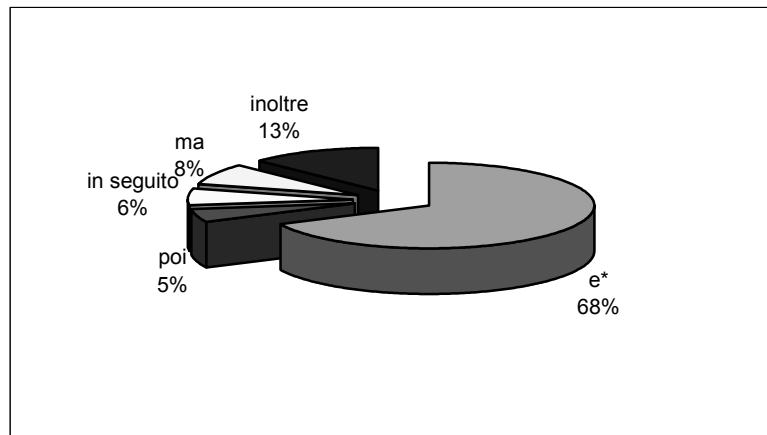
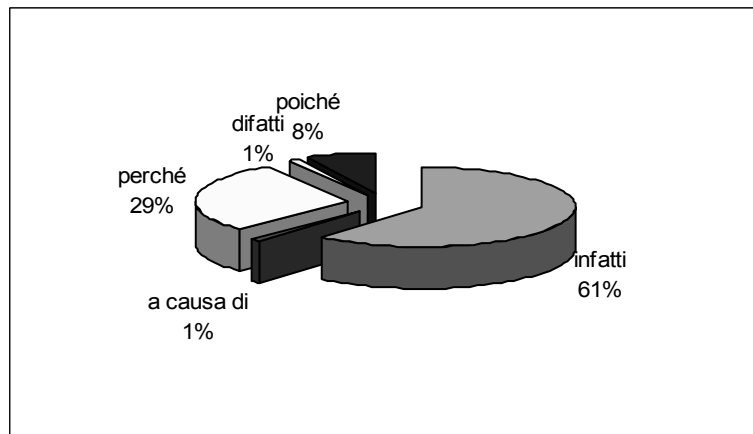
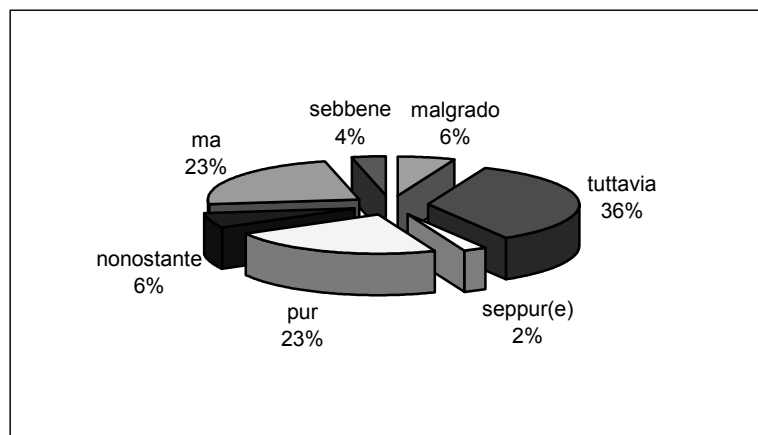
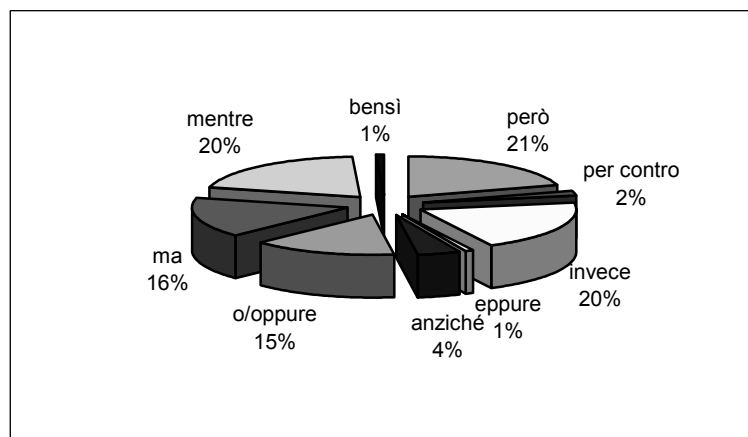
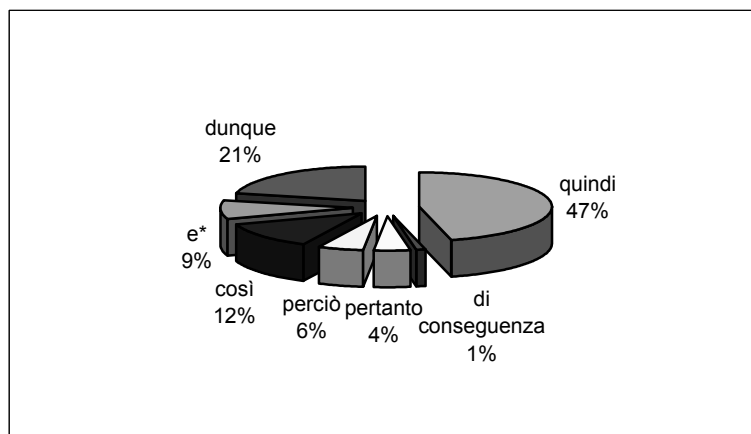
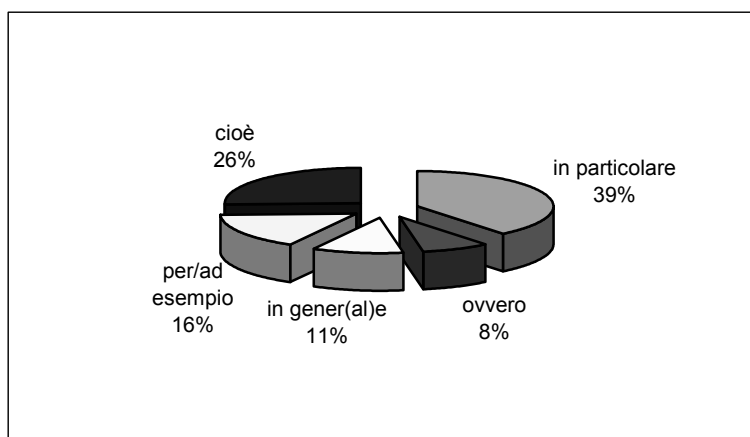
si constata anzitutto che il 10% dei legami di consecuzione è “travestito” da connessione di aggiunta tramite la congiunzione *e* [5], come nel caso di:

[6] `Questa colorazione non è uniforme` **`e`** `ne risulta un insieme di tinte armoniche dove l’ uso del colore risulta funzionale per evidenziare le parti in rilievo e alcune parti decorative.`

Athenaeum.

A questa osservazione va poi aggiunto che, come si è detto in Ferrari 2005b, pp. 271-272, in molti casi in cui ci potrebbe essere potenzialmente un chiaro movimento consecutivo si preferisce optare per segnali linguistici con una semantica ambigua tra la consecuzione e la motivazione (...: *tutti segni che*...), per una gerundiale post-reggente, per una subordinazione relativa. Si tratta, complessivamente, di fenomeni che in un certo senso snaturano il dinamismo testuale insito nella consecuzione, in quanto la calano in configurazioni che collocano il contenuto consecutivo sulla stesso piano o su un piano informativo inferiore rispetto a quello in cui si inserisce la premessa. Lo stesso tipo di ragionamento può essere applicato alla classe delle relazioni logiche di rielaborazione. Anche in questo caso si osservano, *mutatis mutandis*, i fenomeni linguistici rilevati per la consecuzione, a cui va aggiunta un altro dato significativo. L’esemplificazione e la riformulazione parafrastica compaiono volentieri racchiuse tra parentesi. Ora, se la specificità comunicativa della manifestazione “parentetica” consiste nel creare un piano testuale esterno e di importanza secondaria rispetto al piano semantico centrale del testo (Cignetti 2004), la scelta delle parentesi conferma in altro modo la generale strategia di evitamento delle relazione rielaborative.

[5] I dati quantitativi relativi alla congiunzione *e* fanno riferimento, per evidenti ragioni “pratiche”, a una ricerca svolta su un campione di 100 occorrenze. Di qui l’asterisco, accanto alla congiunzione, nelle tavole 2, 3 e 7.

3.2 La distribuzione dei connettivi. La distribuzione dei vari connettivi all'interno di ogni classe relazionale mostra, come indicano i dati proposti qui di seguito, un tasso relativamente basso di variazione (eccetto forse per i connettivi di contrasto):

Tav. 3: I connettivi di aggiunta.

Tav. 4: I connettivi di motivazione.

Tav. 5: I connettivi di concessione e di limitazione.

Tav. 6: I connettivi di contrasto.

Tav. 7: I connettivi di consecuzione.

Tav. 8: I connettivi di rielaborazione.

Il fenomeno è anzitutto significativo dal punto di vista del registro linguistico che caratterizza i nostri notiziari, il quale, non perseguendo la *variatio* lessicale, si scosta (almeno per questo aspetto) dallo standard-letterario per avvicinarsi piuttosto ad uno stile "medio". L'uniformità dei connettivi è poi interpretabile anche in chiave semantico-pragmatica. Poiché, come mostrano gli studi degli ultimi decenni (si pensi in area francofona ai lavori proposti nell'àmbito dell'*analyse du discours* dall'équipe ginevrina di Eddy Roulet – cfr. Roulet et alii 1985 e 2001 – ed ai lavori di Corinne Rossari), la variazione delle forme dei connettivi porta con sé anche importanti modulazioni semantiche, la monotonia dei connettivi è sintomatica di una certa "piattezza concettuale" delle relazioni logiche, che si ripetono sempre uguali a se stesse. Una piattezza che, a ben guardare, si colora di genericità: i connettivi più frequenti all'interno di ogni singola classe sono tipicamente quelli semanticamente più poveri.

I dati proposti nelle Tavv. da (3) a (8) e la loro interpretazione stilistico-semantica avvicinano tipologicamente i notiziari accademici di *L'Ateneo* alla macroclasse dei testi tecnico-scientifici, in cui il tratto della precisione e della ricchezza, più che nell'architettura del discorso, si colloca nell'àmbito del valore semantico-denotativo di ogni singola proposizione. Altri aspetti della loro architettura logica ne fanno tuttavia dei testi tecnico-scientifici *sui generis*, in cui l'ampiezza e la trasparenza del movimento esplicativo è sostituita da un andamento giustappositivo tipico, appunto, del notiziario.

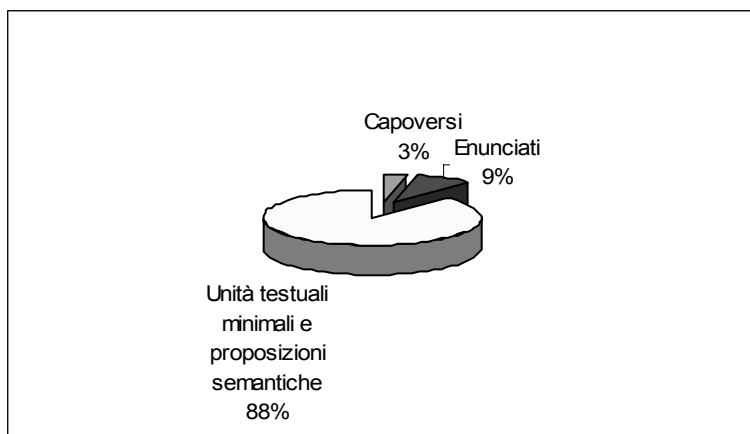
4. I LIVELLI TESTUALI DELLE ARTICOLAZIONI LOGICHE. Un'altra questione è come si intesechino connettivi e livelli testuali delle articolazioni logiche nei notiziari accademici.

4.1 I "LUOGHI" DELLE RELAZIONI LOGICHE. Un aspetto cruciale, e tuttavia per lo più trascurato, della caratterizzazione dell'architettura dei testi consiste nell'identificazione dei "luoghi" in cui si concentrano le relazioni logiche. In astratto ed in generale, esse infatti non sono specializzate per un livello particolare, ma attraversano il testo in tutti i suoi spazi potendo interessare tutte le sue unità costitutive, da quelle più piccole a quelle di ordine più elevato. Così (cfr. per un approfondimento Ferrari 2005a e 2005b), le possiamo trovare tra proposizioni semantiche, tra unità informative – cioè tra contenuti semantico-pragmatici unitari dal punto di vista della loro funzione informativa –, tra enunciati – *i.e.* contenuti semantico-pragmatici caratterizzati da unità illocutiva –, tra gruppi di enunciati, tra capoversi, tra paragrafi, tra capitoli ecc.

Per esempio, nel testo [3] che riproduciamo ancora qui di seguito,

[3] In quest'ultimo caso, **vale a dire** nella situazione in cui si trova l'italiano dal Cinquecento ad oggi, possiamo ulteriormente riconoscere dei a) *periodi a normazione rigida*, e dei b) *periodi a normazione debole*. Per fare qualche esempio, possiamo dire che tutto il XVI secolo è stato un periodo a normazione rigida, almeno a partire dal 1525, anno di pubblicazione delle *Prose della volgar lingua* del Bembo, l'atto di fondazione della lingua letteraria comune; **così come** sono stati gli ultimi trenta anni del XIX secolo, **dopo** l'accettazione su scala nazionale della riforma linguistica manzoniana **e** il proliferare di grammatiche ispirate ad essa. Il Novecento, **al contrario**, è un secolo a normazione debole: lo dimostra la scarsa produzione di grammatiche normative, **in rapporto con** la relativa stabilità dello standard linguistico a livello di strutture fonetiche e morfologiche, **mentre** qualche innovazione di rilievo si ha nella sintassi e soprattutto nell'incremento del lessico neologico. Tesi 2001, pp. 8-9,

limitandoci ai punti in rilievo, osserviamo una relazione di contrasto (*al contrario*) tra due sequenze di enunciati; una relazione di esemplificazione (*Per fare qualche esempio*) tra un enunciato ed una coppia di enunciati; relazioni tra unità informative: di specificazione tra il Quadro *In quest'ultimo caso* e l'Appendice *nella situazione in cui si trova l'italiano dal Cinquecento ad oggi*, di susseguenza temporale segnalata da *dopo*, di contrasto attraverso *mentre*.

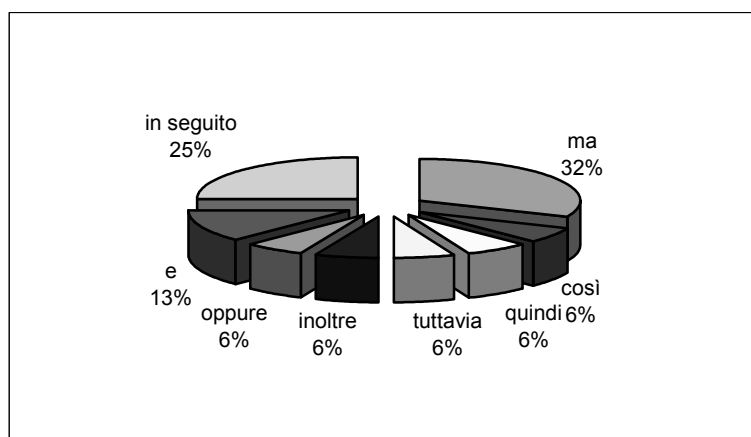
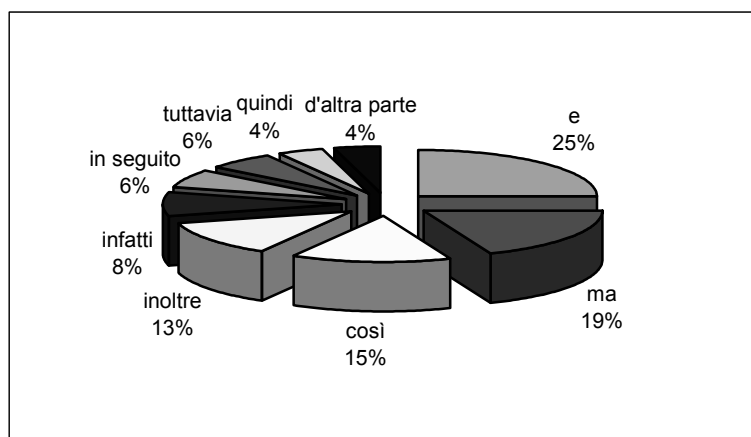
4.2 I "LIVELLI" DEL TESTO. In Ferrari 2005b si osservava che le relazioni logiche dei notiziari accademici coinvolgono essenzialmente i livelli più "bassi" del testo, vale a dire le unità di natura informativa interne all'enunciato e le proposizioni semantiche. Il dato viene confermato in modo ancora più acuto da una ricerca sulla distribuzione dei connettivi entro il corpus, che ha dato i seguenti risultati:

Tav. 9: La distribuzione dei connettivi nei diversi livelli del testo.

Dalla tavola (9) risulta infatti una forte concentrazione dei connettivi pragmatici all'interno dell'enunciato, ed una loro scarsa presenza a cavallo di due capoversi o di due enunciati. Anche quando è espressa dai connettivi, la movimentazione logica dei notiziari accademici si gioca insomma essenzialmente tra le unità minimali del testo e tra le proposizioni semantiche. Il dato è significativo da più punti di vista. Esso mostra anzitutto che la classe di testi in esame privilegia una macro-organizzazione testuale fondata soprattutto su connessioni di tipo tematico: a livello di capoverso e, per una buona parte, anche di enunciato, il discorso non risponde cioè ad un piano organizzativo di tipo logico – con macro-movimenti consecutivi, esplicativi o concessivi – ma ad aggiunte di nuclei semantici la cui connessione è "semplicemente" tematica. Laddove – come dall'enunciato in su – vengono coinvolti valori di tipo illocutivo, che riguardano cioè il

fondamento interattivo stesso dell'atto di dire, la componente logica si scioglie per lasciar posto a quella tematica. Si tratta di un modo di procedere in cui il far sapere vince sulla spiegazione, vale a dire di una modalità di macro-costruzione testuale caratteristica dei discorsi di natura divulgativa e didattica (per un approfondimento, cfr. Ferrari 2005b). Ma c'è di più. Il carattere "discreto" dell'organizzazione logica del testo non è dato solo dal fatto che le relazioni logiche si concentrino all'interno di un singolo atto illocutivo (e non tra atti illocutivi), ma anche dal fatto che esse coinvolgano tipicamente unità semantiche collocate sullo sfondo informativo dell'enunciato: cioè unità che precedono l'informazione principale dell'enunciato o sono inserite al suo interno, come in:

[7] L' aspetto più negativo di questa drammatica e totale distruzione è oggi rappresentato dalla diaspora dei docenti somali , che furono costretti , **malgrado** un iniziale attaccamento alle Università italiane , mantenuto attraverso una rete di rapporti personali con i vecchi professori , a disperdersi in vari Paesi europei ed extraeuropei . Athenaeum.

È questa una proprietà che caratterizza i notiziari accademici, distinguendoli così dai testi genuinamente didattici o divulgativi, come discorsi in un certo senso "specialistici", le cui complessità logiche sono più "suggerite", evocate che non asserite e debitamente sviluppate.

4.3 LA RELAZIONE DI AGGIUNTA. Quanto abbiamo detto nel paragrafo 5.2 viene confermato per altra via dall'osservazione della presenza e distribuzione della relazione di aggiunta. La predilezione dei notiziari accademici per i connettivi di aggiunta, già osservata in Ferrari 2005b, risulta in modo chiaro dalla tavola (1), che attesta che tali connettivi corrispondono al 32% di tutti i connettivi del corpus. La percentuale si fa ancora più elevata quando si guardi ai soli livelli alti del testo:

Tav. 10: Connettivi ad inizio di capoverso.

Tav. 11: Connettivi ad inizio di enunciato.

Come si noterà, sia ad inizio di enunciato che di capoverso la somma delle percentuali dei connettivi di aggiunta *e*, *inoltre*, ed *in seguito* è pari al 42%: ed è una percentuale notevole[6], dal momento che i connettivi non esauriscono le possibilità linguistiche di introdurre una relazione di aggiunta (si pensi ad esempio all'uso dell'avverbio *anche* come introduttore di una relazione aggiuntiva, od alla formula *non solo...ma anche*, entrambi attestati in Ferrari 2005b). La diffusa gestione tematica ed aggiuntiva della connessione tra enunciati, caratteristica delle tipologie espositivo-esplicative, è dunque visibile sia *ex negativo*, via la bassa frequenza di connettivi ad inizio di capoverso e di enunciato, sia in positivo, dalla forte concentrazione di connettivi di aggiunta ai livelli alti del testo.

Un'indagine che tenga conto dei livelli in cui si manifestano le relazioni logiche in *L'Ateneo* permette inoltre di raffinare i dati relativi alla distribuzione "complessiva" dei connettivi nel corpus (Tav. (1)). Un confronto tra le tavole (1), (10) e (11) mette infatti in luce ancora più nitida la concentrazione dell'impianto logico dentro l'enunciato. Così, stando alla tavola (1) – astraendo dunque dal formato delle unità coivolte nella relazione – si osserva che ben il 20% dei connettivi sono di motivazione. Se invece si osservano le tavole (10) e (11), si noterà che soltanto nell'8% dei casi troviamo *infatti* a inizio di enunciato, e mai a inizio di capoverso. Questo significa che la netta maggioranza dei connettivi di motivazione relaziona unità minimali interne all'enunciato o proposizioni semantiche. Se si vuole andare più in là, le tavole (10) e (11) mostrano *ex negativo* che la congiunzione subordinante *perché* (secondo connettivo di motivazione più usato, come rivela la Tav. (4)) non viene mai utilizzata ad inizio di capoverso o di enunciato: il che significa che essa non sceglie mai come primo termine una funzione illocutiva.

Come suggerivamo precedentemente, l'importanza dell'analisi del formato delle unità coinvolte nell'assetto logico-relazionale del testo si misura anche a livello dell'interpretazione semantica dei connettivi: mutando le unità di significato coinvolte (capoversi, enunciati, unità informative, proposizioni semantiche), può in effetti cambiare anche il tipo di relazione veicolata, od il sotto-tipo della sua realizzazione (cfr. le riflessioni in Ferrari 2005b). Per la nostra analisi, è interessante in particolare osservare l'alta frequenza dei connettivi *ma* e (in (11)) *così*: connettivi che, se inaugurali di enunciato, vedono indebolita la loro componente logica. Relativamente alla congiunzione *ma* nella rivista *L'Ateneo*, in Ferrari 2005b si constatava in effetti, sulla linea dei lavori di Marconi e Bertinetto 1984 e di Sabatini 1997, l'affievolirsi della componente av-

[6] Si badi tra l'altro che le tavole (10) e (11) non tengono conto dei casi in cui *inoltre* e *in seguito* non sono incipitari e tuttavia legano enunciati o capoversi.

versativo-limitativa e la funzione sostanzialmente aggiuntiva o di scarto tematico del *ma* incipitario. E l'influsso del formato è riscontrabile anche sulla semantica del connettivo *così*, che perde in parte la sua traccia di consecutività nei casi in cui si lega ad una funzione illocutiva. L'alta percentuale, nelle tavole (10) e (11), delle congiunzioni *ma* e (in (11)) di *così* non fa allora che rafforzare l'idea di una diffusa gestione tematica e aggiuntiva della connessione tra enunciati e capoversi.

5. CONCLUSIONI. L'analisi potrebbe, e dovrebbe, essere ampliata, per esempio attraverso una valutazione attenta dell'intreccio tra significati espliciti e significati impliciti convocato nel testo da ogni classe semantico-concettuale di connettivi. Quanto abbiamo detto nei paragrafi precedenti ci pare tuttavia sufficiente per mostrare quanto possa essere significativa per una caratterizzazione tipologica dei testi la descrizione della loro architettura logico-semantica e una valutazione attenta dell'insieme di connettivi che li caratterizza.

Da un punto di vista metodologico, questa stessa analisi mostra quanto per una ricerca di questo tipo il dato quantitativo – la cui definizione è resa possibile da *corpora* elettronici etichettati – sia nel contempo necessario ma non sufficiente. È necessario in quanto permette di superare – confermandole ma soprattutto modulandole – le speculazioni, o generalizzazioni, impressionistiche che accompagnano tanta letteratura sulla tipologia testuale e di "vedere" fenomeni a cui microanalisi puntuali non permettono di accedere. È insufficiente perché, per cogliere in profondità il senso dell'uso dei connettivi all'interno di un testo, i dati quantitativi devono essere sottoposti ad una serrata valutazione sistematica di carattere qualitativo. Il che significa (j) effettuare, dove la quantità lo riveli pertinente, analisi semantico-lessicali puntuali e profonde, (ij) analizzare attentamente la manifestazione linguistica dei connettivi, prestando attenzione al loro intorno sintattico e interpuntivo, (iij) ragionare all'interno di un solido sistema analitico, senza cui il dato quantitativo e qualitativo non può assumere alcun significato.

BIBLIOGRAFIA[7].

AA. VV.

1997 *Norma e lingua in Italia: alcune riflessioni fra passato e presente. 16 maggio 1996*, Milano, Istituto lombardo di scienze e lettere, 1997 "Incontro di studio" 10.

BEGUELIN - AVANZI

i.p. *Actes du Colloque La Parataxe: Premier Colloque de Macrosyntaxe. Neuchâtel, 12-15 février 2007*, édité par Marie-José Beguelin et Mathieu Avanzi, in preparazione.

BERRENDONNER

1990 Alain Berrendonner, *Pour une macro-syntaxe*, in "Travaux linguistiques" XXI (1990) 25-36.

BONINI - MAZZOLENI

1988 *Linguistica e traduzione: Atti del seminario di studi, Premeno (Novara), Villa Bernocchi, 25-27 settembre 1987*, a cura di Vincenzo Bonini e Marco Mazzoleni, Milano, Comune di Milano, 1988.

CIGNETTI

2004 Luca Cignetti, *Le parentesi tonde, un segno pragmatico di eterogeneità enunciativa*, in FERRARI 2004, pp. 165-189.

[7] La bibliografia proposta è più ampia di quella esplicitamente indicata nell'articolo: ci è sembrato importante fornire le indicazioni che inquadrano in generale le nostre riflessioni in corso sulla semantica dei connettivi.

COLLODI
1981 Carlo Collodi, *Le avventure di Pinocchio*, Milano, Arnoldo Mondadori, [1883] 1981.

CORINO - MARELLO - ONESTI
2006 *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 | Proceedings of the XII EURALEX International Congress. Torino, Italia, 6th-9th September 2006*, a cura di Elisa Corino, Carla Marello e Cristina Onesti, 2 voll., Alessandria, Edizioni dell'Orso, 2006.

CORNULIER
1985 Benoît de Cornulier, *Effets de sens*, Paris, Éditions de Minuit, 1985.
1985a Benoît de Cornulier, *Treize* si *à la douzaine*, in CORNULIER 1985, pp. 56-93.

CRESTI
2000/I Emanuela Cresti, *Corpus di italiano parlato*. *Introduzione*, Firenze, Accademia della Crusca, 2000.
2000/II Emanulea Cresti, *Corpus di italiano parlato*. *Campioni*, Firenze, Accademia della Crusca, 2000.
i.p. *Nuove prospettive nello studio del lessico. Atti del IX Congresso Internazionale SILFI. Firenze, 15-17 Giugno 2006*, a cura di Emanuela Cresti, in preparazione.

DE CESARE
2004 Anna-Maria De Cesare, *L'avverbio* anche *e il rilievo informativo del testo*, in FERRARI 2004, pp. 191-218.

DUCROT
1983 Oswald Ducrot, *Opérateurs argumentatifs et visée argumentative*, in "Cahiers de Linguistique Française" V (1983) 7-36.

FAVA
1995 Elisabetta Fava, *Tipi di atti e tipi di frasi*, in *GGIC* III, pp. 19-48.

FERRARI
1995 Angela Ferrari, *Connessioni. Uno studio integrato della subordinazione avverbiale*, Genève, Slatkine, 1995.
1999 Angela Ferrari, *Tra rappresentazione e esecuzione: indicare la 'causalità testuale' con i nomi e con i verbi*, in "Studi di grammatica italiana" XVIII (1999) 113-144.
2003 Angela Ferrari, *Le ragioni del testo. Aspetti morfosintattici e interpuntivi dell'italiano contemporaneo*, Firenze, Accademia della Crusca, 2003.
2004 *La lingua nel testo, il testo nella lingua*, a cura di Angela Ferrari, Torino, Istituto dell'Atlante Linguistico Italiano, 2004 "Bollettino dell'Atlante linguistico italiano. Supplementi" 9.
2004a Angela Ferrari, *Le subordinate causali nell'architettura del testo*, in FERRARI 2004, pp. 43-78.
2005 *Rilievi. Le gerarchie semantico-pragmatiche di alcuni tipi di testo*, a cura di Angela Ferrari, Firenze, Franco Cesati Editore, 2005 "Quaderni della rassegna" 44.
2005a Angela Ferrari, *Tipi di testo e tipi di gerarchie testuali, con particolare attenzione alla distinzione tra scritto e parlato*, in FERRARI 2005, pp. 15-51.
2005b Angela Ferrari, *Le trame "logiche" dei notiziari accademici*, in FERRARI 2005, pp. 247-292.
2006 *Parole frasi testi tra scritto e parlato*, a cura di Angela Ferrari, Lugano, Cenobio Edizioni, 2006 = "Cenobio" LV (2006)[3] n.s.
2006a Angela Ferrari, *Alternative riformulative*, in CORINO - MARELLO - ONESTI 2006, pp. 1153-1164.

2006b Angela Ferrari, *La fonction textuelle d'Appendice. De la dislocation à l'apposition, à travers la dimension informationnelle*, in "Cahiers Ferdinand de Saussure" LIX (2006) 55-86.

i.p. Angela Ferrari, *Congiunzioni frasali, congiunzione testuali e preposizioni: stessa logica, diversa testualità*, in CRESTI *i.p.*

FERRARI - ROSSARI

1994 Angela Ferrari - Corinne Rossari, *De* donc *à* dunque *e* quindi: *les connexions par raisonnement inférentiel*, in "Cahiers de linguistique française" XV (1994) 7-49.

FERRARI - MANDELLI

i.p. Angela Ferrari - Magda Mandelli, *Virgules, et coordination: aspects sémantiques, informationnels, et textuels*, in BEGUELIN - AVANZI *i.p.*

FISCHER

i.s. *Approaches to Discourse Particles*, edited by Kerstin Fischer, Amsterdam, Elsevier, in corso di stampa.

FRASER

1999 Bruce Fraser, *What are Discourse Markers?*, in "Journal of Pragmatics" XXI (1999) 931-952.

GGIC I → RENZI - SALVI et alii 1988; II → RENZI - SALVI et alii 1991; III → RENZI - SALVI et alii 1995.

HALLIDAY

1985 Michael A[lexander] K[irkwood] Halliday, *An Introduction to Functional Grammar*, London - Baltimore, Edward Arnold, 1985.

KNOTT - SANDERS

1998 Alistair Knott - Ted Sanders, *The Classification of Coherence Relations and their Linguistic Markers: An Exploration of two Languages*, in "Journal of Pragmatics" XXX (1998) 135-175.

LALA

2004 Letizia Lala, *I Due punti e l'organizzazione logico-argomentativa del testo*, in FERRARI 2004, pp. 143-164.

MANDELLI

2004 Magda Mandelli, *Coordinazione frasale e coordinazioni testuali: il caso della congiunzione* e, in FERRARI 2004, pp. 117-142.

i.p. Magda Mandelli, In effetti *nel testo*, in CRESTI *i.p.*.

MANZOTTI

1987 Emilio Manzotti, *I costrutti cosiddetti eccettuativi in italiano, inglese e tedesco: semantica e pragmatica*, in BONINI - MAZZOLENI, pp. 67-110.

MARCONI - BERTINETTO

1984 Diego Marconi - Pier Marco Bertinetto, *Analisi di* ma. *Parte prima: semantica e pragmatica*, in "Lingua e stile" XIX (1984)[2] 223-258.

PASCH - BRAUßE - BREINDL

2003 Renate Pasch - Ursula Brauße - Eva Breindl, *Handbuch der deutschen Konnektoren*, Berlin - New York, Walter de Gruyter, 2003 "Schriften des Instituts für deutsche Sprache".

ROSSARI
1993 Corinne Rossari, *Les opérations de reformulation: analyse du processus et des marques dans une perspective contrastive français-italien*, Bern - Berlin - Paris, B. Lang, 1993.
2000 Corinne Rossari, *Connecteurs et relations de discours: des liens entre cognition et signification*, Nancy, Presses Universitaires de Nancy, 2000.
i.s. Corinne Rossari, *The formal properties of a subset of discourse markers: connectives*, in FISCHER *i.s.*

ROULET
2001 Eddy Roulet, *L'organisation relationnelle*, in ROULET et alii 2001, pp. 165-199.

ROULET et alii
1985 Eddy Roulet - A[ntoine] Auchlin, J[acques] Moeschler, C[hristian] Rubattel, M[arianne] Schelling, *L'articulation du discours en français contemporain*, Berne - Frankfurt a.M. - New York, P. Lang, 1985.
2001 Eddy Roulet - Laurent Filliettaz - Anne Grobet, *Un modèle et un instrument d'analyse de l'organisation du discours*, Bern - Berlin - Bruxelles, Peter Lang, 2001.

SABATINI
1997 Francesco Sabatini, *Pause e congiunzioni nel testo. Quel* ma *a inizio di frase…*, in AA. VV. 1997, pp. 113-46.

SBISÀ
1978 *Gli atti linguistici. Aspetti e problemi di filosofia del linguaggio*, a cura di Marina Sbisà, Milano, Feltrinelli, 1978.
1989 Marina Sbisà, *Linguaggio, ragione, interazione. Per una teoria pragmatica degli atti linguistici,* Bologna, Il Mulino, 1989.

SCARANO
2002 Antonietta Scarano, *Frasi relative e pseudo-relative in italiano. Sintassi, semantica e articolazione dell'informazione*, Roma, Bulzoni, 2002.

SCHIFFRIN
1987 Deborah Schiffrin, *Discourse Markers*, Cambridge, Cambridge University Press, 1987 "Studies in Interactional Sociolinguistics" 5.

SPERBER - WILSON
1986 Dan Sperber - Deirdre Wilson, *Relevance. Communication and Cognition*, Cambridge, Harvard University Press, 1986 "Language and Thought Series".

TESI
2001 Riccardo Tesi, *Storia dell'italiano: la formazione della lingua comune dalle origini al Rinascimento*, Roma-Bari, Laterza, 2001.

VISCONTI
2000 Jacqueline Visconti, *I connettivi condizionali complessi in italiano e in inglese. Uno studio contrastivo*, Alessandria, Edizioni dell'Orso, 2000 Gli argomenti umani" 5.

ZAMPESE
2004 Luciano Zampese, *Aspetti semantico-testuali del gerundio modale in apertura di frase*, in FERRARI 2004, pp. 79-116.

CORPORA DI RIFERIMENTO.

Athenaeum Corpus `http://www.bmanuel.org/projects/at-HOME.html`.

LUCA CIGNETTI
(Università di Basilea)

11. Alcune forme di polifonia testuale nei notiziari accademici di *Athenaeum*. *Aspetti funzionali ed argomentativi.*

0. INTRODUZIONE. L'uso di "polifonia" come termine specialistico risale agli studi di àmbito stilistico-letterario di Mixail Baxtin, dove è usato per indicare l'intreccio di voci caratteristico dello stile di Dostoevskij e costituitivo, secondo l'autore, del romanzo moderno[1]. L'interpretazione più fertile, in seno alle scienze del linguaggio, fu in séguito formulata da Oswald Ducrot (cfr. Ducrot et alii 1980, ecc.), che definisce "polifonico" l'enunciato in cui compare una pluralità di voci, ma non necessariamente una pluralità di locutori. La polifonia così intesa implica una pluralità di punti di vista, introdotti nel testo per ottenere particolari fini argomentativi e soprattutto disposti in rapporto di tipo gerarchico l'uno rispetto all'altro.

Naturalmente, un'interpretazione così ampia non può che avere manifestazioni linguistiche molto diverse; tra i molti fenomeni che possono essere descritti facendo ricorso a questo paradigma (cfr. Nølke 1994) il più esplicito è forse il discorso riportato (DR), inteso come l'enunciato che viene prodotto in un atto di enunciazione diverso da quello di cui la citazione fa parte[2]. Il discorso riportato è in genere introdotto da segnali linguistico-testuali espliciti, detti "introduttori locutivi" (cfr. Cresti 2000), la cui presenza è molto più diffusa nel parlato: nel corpus *C-Oral Rom*, composto da 310.969 parole, il lemma *dire* (l'introduttore locutivo più tipico) compare 3.234 volte, pari al 10,4 ‰ delle parole totali (è il terzo verbo per frequenza, dopo *essere* e *fare*). Se, con buona approssimazione e per soli fini statistici, consideriamo la presenza del lemma *dire* come indicatore della frequenza di DR, osserviamo che nel corpus *Athenaeum* questa forma di polifonia è sottorappresentata, anche rispetto ad altri testi scritti di tipo funzionale. Nel corpus di italiano funzionale *LISULB*[3], ad esempio, l'introduttore locutivo *dire* compare 1.549 volte, pari al 1,3 ‰ delle parole totali, mentre nel sottocorpus di *Athenaeum* preso in analisi[4], composto da 55.589 parole, compare solo in 9 casi, pari allo 0,16 ‰ del totale:

---

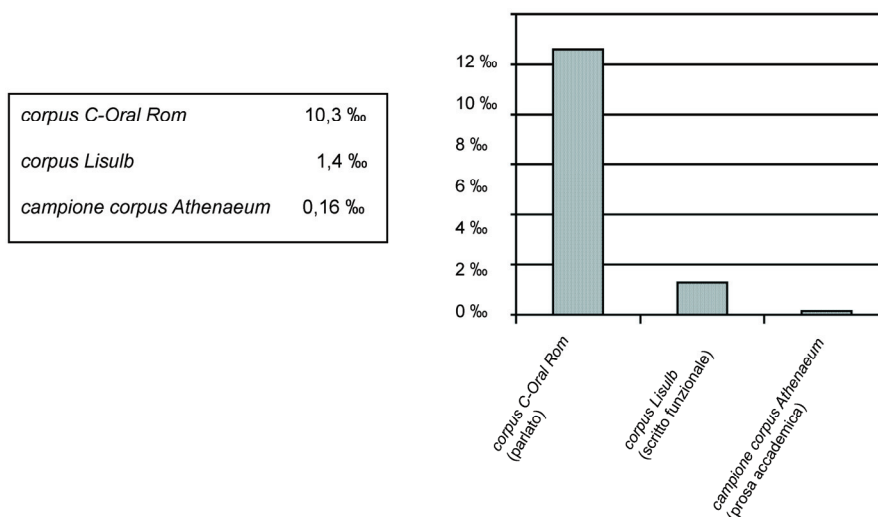
[1] Cfr. Baxtin 1970/29, p. 35: «Dostoevskij est le créateur du roman polyphonique. Il a élaboré un genre romanesque nouveau». Cfr. anche *ibid.*: «Ce qui apparaît dans ses œuvres ce n'est pas la multiplicité de caractères et de destins, à l'intérieur d'un monde unique et objectif, éclairé par la seule conscience de l'auteur, mais la pluralité des consciences "équipollentes" et de leur univers qui, sans fusionner, se combinent dans l'unité d'un événement donné. Les héros principaux de Dostoevskij sont, en effet, dans la conception même de l'artiste, non seulement objets de discours de l'auteur, mais sujets de leur propre discours immédiatement signifiant».

[2] I tipi classici di DR sono il discorso diretto (DD), caratterizzato dalla presenza di due locutori, di due contesti deittico-situazionali e di due tempi di riferimento (tipico del DD è inoltre la marca della parte citata per mezzo di virgolette, lineette o caratteri in corsivo); il discorso indiretto (DI), dove l'enunciato citato è integrato nell'enunciato citante, con l'effetto di una pluralità locuzionale collocata in un sistema di riferimento deittico-temporale univoco (come marca del segmento citato si ha perlopiù *verba dicendi* seguiti dal connettivo *che*, da proforme con pari funzione, da interrogative dirette o dal *di* con infinito); ed il discorso indiretto libero (DIL), dove i due locutori sono posti in due contesti deittico-situazionali ma in un solo tempo di riferimento, corrispondente a quello del locutore citante (l'uso di questo tipo è limitato, in genere, ai testi letterari). Cfr., in merito, Mortara Garavelli 1985.

[3] Il *corpus LISULB* (Linguistica Italiana Sincronica Università di Losanna e Università di Basilea) è composto da estratti di lingua scritta funzionale (non letteraria) di varia tipologia: saggistica letteraria, saggistica linguistica, quotidiani e riviste, testi giuridici e manuali didattici, per un totale di 1.225.830 parole.

[4] Il canone è così composto: *Energia e ambiente: una nuova politica per l'ambiente; Indagine del C.I.R.D.A sulla ricerca didattica nell'Università di Torino: aspetti quantitativi e qualitativi; Per una città capace di futuro; Adempimenti legislativi per la tutela del benessere animale e della protezione dei lavoratori negli stabulari*

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp.199-207.

corpus C-Oral Rom	10,3 ‰
corpus Lisulb	1,4 ‰
campione corpus Athenaeum	0,16 ‰

Tav. 1: Occorrenze del lemma *dire*

1. UNA PROSA "MONOFONICA"? Come osservato in Ferrari 2005b (pp. 245 sgg.), la prosa accademica raccolta nel corpus *Athenaeum* non corrisponde ad un tipo testuale omogeneo, ma a testi diversi per funzioni retorico-illocutive e per il rapporto che essi intrattengono con l'orale. Si è detto che le forme di discorso riportato in questo corpus sono sottorappresentate rispetto ad altre tipologie di testo informativo-esplicativo. Quando compaiono, la loro funzione è in genere di tipo argomentativo-esornativo, soprattutto nella forma del DD, come in [1]:

[1] `Il Palazzo resta , con la chiesa di San Francesco da Paola , una delle tappe essenziali nella via Po , un' arteria che indirizza alla piazza Castello , in un percorso tanto apprezzato da Nietzsche ( 1888 ) :` **`" scorgere le Alpi dal centro della città ! Queste lunghe strade che sembrano condurre in linea retta verso le auguste cime nevose . Aria serena , limpida in modo sublime . Non avrei mai creduto che una città , grazie alla luce , potesse diventare così bella … Si può camminare per mezze ore di seguito sotto alti portici . Qui tutto è costruito con liberalità ed ampiezza , specialmente le piazze , così anche nel cuore della città si ha un senso superbo di libertà " .`** Athenaeum.

Nell'esempio riportato, la voce citata gode di un alto prestigio intellettuale, funzionale al contesto accademico, ed esercita una funzione esornativa ed aulicizzante; altrimenti, la fonte citata può essere anche la fonte delle informazioni, come in [2]:

dell'Università di Torino; Ricerche etnologiche e accordi di cooperazione dall'Africa Equatoriale all'Africa Occidentale; L'Africa e il centro per lo studio delle Letterature e delle culture delle aree emergenti; L'Università di Torino e l'Africa letteraria di espressione francese; Attività della missione archeologica italiana a Abuqir (Egitto); Gli scavi archeologici di Abuqir; Sostenibilità ambientale, sostenibilità umana. Alcune esperienze africane di ricerca-azione; La facoltà di medicina veterinaria e l'Africa; L'attività del CNR nel settore amianto tra passato e futuro; Epidemiologia delle malattie da amianto in Italia; La rifondazione dello Studio torinese: Vittorio Amedeo II e l'Università; Il settecentesco Palazzo degli Studi; Il cantiere di restauro; Il restauro del Palazzo dell'Università di Torino; L'analisi dei trattamenti murali negli stucchi e decorazioni murarie; La ricerca storica quale strumento finalizzato al restauro; La ricerca nei laboratori scientifici "A. Mosso"; Proposta di recupero dell'istituto scientifico "A. Mosso"; Animali africani nel museo di zoologia dell'Università di Torino; La mostra "L'Africa in Piemonte tra '800 e '900"; Recenti sviluppi della "questione amianti in Italia".

[2] Il viaggio del Donati fu assai avventuroso e drammatico , almeno a quanto ci tramanda Michele Lessona (1877) : " **Sopra una nave turchesca trabalzata dai flutti tempestosi dell' oceano indiano , ora è poco più di un secolo , agonizzava un uomo partito da Torino nel meglio della vita col proposito generoso di arricchire la nostra città dei prodotti naturali di lontane regioni " .** Athenaeum

In quest'ultimo caso il DR ha funzione modalizzante, introduce infatti una mitigazione della forza illocutiva attraverso l'attribuzione della responsabilità ad altri[5] ("schermo", nei termini di Caffi 2001, p. 321). Altro sfruttamento è l'uso della citazione a fini compositivi, ad esempio come artificio di incipit:

[3] **" Riflettendo Noi all' avvantaggio che può apportare ai nostri Popoli l' eriger , e stabilire in questa nostra Città un Università che provvista di Maestri , e Lettori in tutte le scienze possa dare conveniente pascolo , et alieno non solo alla Gioventù de nostri Stati , che vorranno accedervi , ma anche a quello de Stati alieni che invitata , potrà introdurvisi , ove tanto gli uni , quanto gli altri saranno instato d' habilitarsi in quelle d' esse scienze , nelle quali avranno maggior propensione , e per riuscirvi comodamente restando indispensabili la Costruzione d' una fabricha non men decorosa , che comoda , e ben capace per alloggiarvi detti Lettori , e Maestri separatamente affinche ogn' uno d' essi possa far le sue funzioni senza incomodo degli altri . A qual effetto habbiamo destinato il sito che si è creduto più proprio per tal costruzione e lasciati i nostri ordini per darvi principio presentemente , in maniera che fra tre anni compreso il corrente sia interamente compita , e resa habitabile ... " .** Così scrive Vittorio Amedeo II il 9 Marzo 1713 , sancendo l' avvio della realizzazione del nuovo Palazzo dell' Università di Torino . Athenaeum.

Nei casi di DI, invece, l'*auctoritas* è in genere uno specialista della materia: il valore argomentativo appare allora più esplicito, in quanto la fonte citata è anche il garante dell'asserzione dell'autore:

[4] Storicamente questo monopolio , facilitato dalle pubblicazioni in una lingua che è di fatto l' esperanto di ogni comunità scientifica e da una solida tradizione di ricerca empirica , ha prodotto una serie di benefici per le ricerche mediologiche . Quello più eclatante riguarda la nuova " visibilità " dell' audience , emancipata dal ruolo di categoria residuale o di simulazione cui gli studi quantitativi e l' approccio finalizzato alla verifica degli effetti l' avevano relegata . **Una visibilità ,** come sostiene Sonia Livingstone **, che è insieme teoretica , empirica e politica ,** se si pensa al forte intento emancipatorio che ha caratterizzato i Cultural Studies fin dalla

[5] Cfr. Mortara Garavelli 1985, p. 56: «Il locutore può enunciare proposizioni sulla cui verità od attendibilità non vuole o non può pronunciarsi, per svariati motivi; e in tali casi egli ha a propria disposizione mezzi sintattici e lessicali (per es. il condizionale, espressioni come: *secondo x...*, *a parere di x...*, *a detta di x...*, *a sentire x*, *a voler credere a...*, ecc., eventualmente rinforzate [...] da espedienti grafici per prendere le distanze da ciò che riporta, per dissociarsi dalla responsabilità delle asserzioni contenute nell'atto di enunciazione; per far capire che egli è solo il *locutore* (colui che enuncia), ma che l'*enunciatore* (il responsabile della verità di ciò che viene asserito) è un altro».

> loro nascita. Non dimentichiamo poi che , attorno all' audience , ruotano questioni cruciali come il rapporto tra individuo e società , tra dimensione micro e macro , tra struttura e azione , tra libertà e determinismo , sia esso testuale o sociale .
>
> Athenaeum.

Benché da un punto di vista funzionale il DR mostri una molteplicità di impieghi, quantitativamente, si è detto, non è adeguatamente rappresentato. Del resto lo sfruttamento naturale del DR, in contesti non letterari, è di tipo argomentativo, e la prosa in oggetto si caratterizza per il «prevalere – caratteristico dei testi cosiddetti "informativi" – della componente espositiva su quella argomentativa» (Ferrari 2005b, p. 270).

2. CORI PER VOCE SOLA. Se la presenza di DR, in tutte le sue forme, è molto limitata nel tipo di testo rappresentato dal corpus *Athenaeum*, relativamente più frequenti sono le costruzioni parentetiche, che, come si vedrà, possono essere considerate fenomeni di polifonia testuale. Ferrari 2005b osserva come in questa prosa le parentesi siano sfruttate per accogliere le relazioni elaborative e "strutturanti"; il tipo testuale rappresentato da *Athenaeum*, inoltre, predilige espressioni connettive semanticamente povere, segno di una tendenziale genericità dell'architettura logica: questo fenomeno caratterizzerebbe in generale i testi espositivi, «in particolare le esposizioni di carattere "tecnico" e scientifico, in cui il tratto della precisione è affidato soprattutto alle forme lessicali di carattere denotativo» (Ferrari 2005b, p. 275). Le costruzioni parentetiche si prestano a quest'uso, poiché permettono il recupero inferenziale anche in assenza di connettivi, soprattutto per le relazioni di tipo "motivazione":

[5] Come la relatività di Einstein si riduce a quella di Galileo per velocità " a misura d' uomo " , piccole rispetto alla velocità della luce , così le predizioni della MQ non differiscono da quelle classiche sulla scala dell' esperienza quotidiana **(il pensiero scientifico si sviluppa senza gettare via nulla : ogni nuova teoria deve soddisfare il " principio di permanenza " , quindi riprodurre i risultati della vecchia laddove questa è in accordo con l' esperienza) .** Athenaeum.

Se osserviamo il brano nell'esempio [5], possiamo riconoscere con facilità la presenza di due enunciati[6], in cui il secondo è posto all'interno del primo, che può tuttavia concludersi anche dopo la sua fine (la demarcazione tra i due livelli enunciativi è infatti costantemente garantita dalle parentesi). In casi come questi le parentesi sono sfruttate come segnali di dicotomia enunciativa e possono combinarsi anche con enunciati di orientamento illocutivo diverso, come nel caso di un'asserzione e di un'esclamazione. Grazie alle caratteristiche delle parentesi, in altri termini, è possibile introdurre più enunciati su piani diversi, senza che la coesione del testo risulti compromessa:

[6] Un uso corretto del supporto documentario permette invece di operare il riconoscimento della facies originaria , così come degli interventi successivi , evitando il rischio dell' interpretazione di quella odierna come veritiera **(quante pagine critiche sono state scritte solo sulla base di una fotografia !)** senza tener conto della patina del tempo , dello opere seguenti (determinate dal fatto che l' architettura è un oggetto vivo che muta nel tempo) e delle rifiniture spesso rimaste solo a livello di intenzione . Athenaeum.

[6] Intesi come i corrispettivi linguistici di un atto illocutivo e di composizione testuale (cfr. Cresti 2000, Ferrari 2003, Cignetti 2004 e Ferrari 2004 e 2005).

In questo senso le parentesi possono essere considerate fattori di polifonia testuale, poiché producono lo sdoppiamento del discorso anche in assenza di quello dei locutori. Si realizza, in altri termini, un testo in cui uno stesso locutore asserisce e commenta un fatto attribuendovi, su diversi piani enunciativi, un valore di verità od accettabilità, oppure una marca di tipo affettivo e/o assiologico, oppure ancora modalizzando l'enunciazione[7]. In questa configurazione "auto-dialogica", il locutore parentetico risulta tendenzialmente orientato verso la propria individualità: ecco che allora la parentesi si configura, per usare le parole di Pétillon-Boucheron (2002, p. 333), come "un lieu en marge, un lieu subjectif". Il brano che segue illustra il caso in cui il locutore interviene dall'esterno per commentare soggettivamente il proprio enunciato:

[7] Alla conferenza nazionale hanno fatto seguito due iniziative rilevanti , la proposta di un progetto strategico da parte del Ministero della Sanità **(purtroppo a tutt' oggi finanziato solo in minima parte)** su " Amianto e materiali sostituivi " e la costituzione presso il CNR di una commissione " Amianto dismissione e sostituzione " , presieduta dalla professoressa Anna Marabini , avente lo scopo di formulare e coordinare progetti di ricerca circa la rimozione , inattivazione e sostituzione dell' amianto . Athenaeum.

Ma l'intervento del locutore può manifestarsi in inciso anche per esplicitare la gerarchia informativa dei dati trasmessi:

[8] Più precisamente egli condusse , dal 1903 al 1914 ogni anno , quindi nel 1921 , una serie di campagne di scavo : ne trasse reperti con i quali arricchì notevolmente il Museo , e inoltre **- ciò che qui più interessa -** scrisse con esse due pagine nuove per la nostra scienza : rivelò la cultura fiorita nell' Egitto antico fuori di Menfi e Tebe , nella provincia , e diede il via agli studi di antropologia fisica appuntati sulla popolazione locale . Athenaeum.

Oppure, l'enunciato tra parentesi può essere sfruttato come strategia argomentativa, ad esempio per prevenire l'obiezione dell'interlocutore ed aggiungere un secondo argomento utile alla validazione della tesi principale. Sono questi i casi in cui l'atto parentetico è illocutivamente "sussidiario" al principale, perché funzionale alla sua realizzazione (cfr. Motsch - Pasch 1987 e Cignetti 2004):

[9] Si portano infatti appresso non soltanto il ricordo di mesi intensissimi sul piano delle sollecitazioni intellettuali e professionali e delle emozioni vissute ma anche molto spesso nuove consapevolezze tanto di ciò che significano in concreto le disparità Nord - Sud quanto delle grandi risorse di abilità e di saperi grazie a cui donne e uomini di paesi ' poveri ' riescono a vivere **(certo assai duramente ma per certi aspetti anche più sanamente di noi : qualche giorno e qualche notte in un villaggio africano espone a percezioni radicali dello scarto che divide lo spreco consumistico dall' essenzialità dei bisogni) .** Athenaeum.

[7] Come fanno ad es. gli avverbi di enunciazione, la cui funzione è di «segnalare l'atteggiamento del parlante verso l'enunciazione» (Conte 1999/88, p. 49).

In altri casi, le parentesi sono sfruttate polifonicamente per ottenere effetti testuali complessi: il brano citato in [10] è posto al centro dell'argomentazione con pieno valore assertivo, cosicché il locutore possa sfruttarne appieno il contenuto sia semantico sia pragmatico per i fini che il proprio testo richiede, pur attribuendone la responsabilità ad altri:

[10] La comparazione con i voti medi di altre realtà universitarie vede in posizione di svantaggio i laureati della facoltà di Economia di Torino che sono distanti di oltre un punto dai colleghi delle facoltà economiche dell' Emilia e Romagna e sono al di sotto di due punti della media nazionale , che si attesta su 102 . Quali le cause della più bassa valutazione dei laureati della facoltà di Economia di Torino ? Non si può , solo ed in modo alquanto superficiale , imputarla **- stando almeno alle opinioni ricorrenti tra gli studenti -** ad una maggiore ristrettezza di valutazione dei docenti anche se non si può del tutto escluderla . Athenaeum.

La possibilità di porre informazioni al di fuori dell'asse centrale del testo attraverso le parentesi rende queste ultime in grado di accogliere relazioni logiche fortemente "strutturanti", come la riformulazione parafrastica, l'illustrazione e l'esemplificazione, poco sfruttate nella prosa accademica (cfr. Ferrari 2005b, pp. 270-271). Altro effetto strutturante – ed evidentemente polifonico – è quello in cui, tra parentesi, sono fornite informazioni di tipo compositivo, ad esempio nel caso in cui è annunciato il contenuto di sezioni future del testo:

[11] Si tratta dunque di un continente con cui l' intreccio dei rapporti su tematiche agro-forestali e ambientali è al momento molto ricco ; si noti , anzi , che la fotografia della situazione in atto nel '98-'99 non dava conto di altre esperienze **(anche molto consistenti : di alcune di esse accenneremo in seguito)** appena concluse o che comunque hanno avuto luogo nel corso del decennio . Athenaeum.

In tutti gli esempi illustrati, la polifonia parentetica produce sempre un effetto gerarchizzante (cfr. Cignetti 2005 e Ferrari 2005a), tramite l'assegnazione di un basso dinamismo comunicativo all'informazione trasmessa: per questa ragione non è possibile, in genere, una ripresa di *topic* né è possibile inaugurare un nuovo movimento argomentativo muovendo da un referente parentetico (con effetti simili a quelli di "schermo topicale", per cui cfr. Caffi 2001, p. 322 sgg.).

BIBLIOGRAFIA.

BAXTIN
1970/29 Mihail Mihailovič Bahtin, *Problèmes de la poétique de Dostoïevskij*, traduit [de la 2e édition russe] par Guy Verret, Lausanne, L'Age de l'Homme, 1970. [Edizione originale: Михаил Михайлович Бахтин, *Проблемы творчества Достоевского*, 1929].
1979 Michail [Michailovič] Bachtin, *Estetica e romanzo*, a cura di Clara Strada Janovic, Torino, Einaudi, 1979 "Einaudi Paperbacks" 107.

BERTINETTO et alii
1995 *Temporal Reference, Aspect and Actionality*, 1. *Semantic and Syntactic Perspectives*, a cura di Pier Marco Bertinetto, Valentina Bianchi, James Higginbotham e Mario Squartini, Torino, Rosenberg & Sellier, 1995.

BORGATO - SALVI
1995 Gianluigi Borgato - Giampaolo Salvi, *Le frasi parentetiche*, in *GGIC* III, pp. 165-174.

CALARESU
2004 Emilia Calaresu, *Testuali parole. La dimensione pragmatica e testuale del discorso riportato*, Milano, FrancoAngeli, 2004.

CAFFI
2001 Claudia Caffi, *La mitigazione. Un approccio pragmatico alla comunicazione nei contesti terapeutici*, Münster, LIT, 2001.

CASTELNOVO - VOGEL
1995 Walter Castelnovo - Roos Vogel, *Reported Speech*, in BERTINETTO et alii 1995, pp. 255-272.

CIGNETTI
2004 Luca Cignetti, *Le parentesi tonde, un segno pragmatico di eterogeneità enunciativa*, in FERRARI 2004, pp. 165-189.
2005 Luca Cignetti, *Sfondi e rilievi testuali nella Costituzione della Repubblica Italiana*, in FERRARI 2005, pp. 85-135.

CONTE
1999/88 Maria-Elisabeth Conte, *Condizioni di coerenza*, Alessandria, Edizioni dell'Orso, 1999. Nuova edizione, con l'aggiunta di due saggi a cura di Bice Mortara Garavelli, di Maria-Elisabeth Conte, *Condizioni di coerenza. Ricerche di linguistica testuale*, Firenze, La Nuova Italia Editrice, 1988 "Pubblicazioni della Facoltà di Lettere e filosofia dell'Università di Pavia" 46.

CRESTI
2000 Emanuela Cresti, *Corpus di italiano parlato*, 2 voll., Firenze, Accademia della Crusca, 2000.

CRESTI - MONEGLIA
2005 *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, a cura di Emanuela Cresti e Massimo Moneglia, 1 vol. con DVD, Amsterdam, Benjamins, 2005.

DUCROT
1972 Oswald Ducrot, *Dire et ne pas dire: principes de sémantique linguistique*, Paris, Hermann, 1972.
1984 Oswald Ducrot, *Le dire et le dit*, Paris, Éditions de Minuit, 1984.

DUCROT et alii
1980 *Les mots du discours*, par Danièle Bourcier, Sylvie Bruxelles, Anne-Marie Diller, Oswald Ducrot [etc.], sous la direction d'Oswald Ducrot, Paris, Éditions de Minuit, 1980.

FAVA
1995 Elisabetta Fava, *Tipi di atti e tipi di frasi*, in *GGIC* III, pp. 19-48.

FERRARI
2003 Angela Ferrari, *Le ragioni del testo. Aspetti morfo-sintattici e interpuntivi dell'italiano contemporaneo*, Firenze, Accademia della Crusca, 2003.
2004 *La lingua nel testo, il testo nella lingua*, a cura di Angela Ferrari, Torino, Istituto dell'Atlante Linguistico Italiano, 2004 "Bollettino dell'Atlante linguistico italiano. Supplementi" 9.
2005 *Rilievi. Le gerarchie semantico-pragmatiche di alcuni tipi di testo*, a cura di Angela Ferrari, Firenze, Cesati, 2005 "Quaderni della rassegna" 44.

2005a Angela Ferrari, *Tipi di testo e tipi di gerarchie testuali, con particolare attenzione alla distinzione tra scritto e parlato*, in FERRARI 2005, pp. 15-51.
2005b Angela Ferrari, *Le trame "logiche" dei notiziari accademici*, in FERRARI 2005, pp. 245-290.

GGIC I → RENZI - SALVI et alii 1988; II → RENZI - SALVI et alii 1991; III → RENZI - SALVI et alii 1995.

KERBRAT-ORECCHIONI
1980 Catherine Kerbrat-Orecchioni, *L'énonciation. De la subjectivité dans le langage*, Paris, A. Colin, 1980.

LAUSBERG
1969 Heinrich Lausberg, *Elementi di retorica*, introduzione all'edizione italiana di Lea Ritter Santini, Bologna, Il Mulino, 1969. [Edizione originale: Heinrich Lausberg, *Elemente der literarischen Rhetorik: eine Einführung für Studierende der klassischen, romanischen, englischen und deutschen Philologie*, 3. durchgesehene Auflage, München, Max Hueber Verlag, 1967; 1. Auflage 1949].

LO CASCIO
1993 Vincenzo Lo Cascio, *Grammatica dell'argomentare: strategie e strutture*, Scandicci, La Nuova Italia, 1993.

MORTARA GARAVELLI
1985 Bice Mortara Garavelli, *La parola d'altri*, Palermo, Sellerio, 1985.
1995a Bice Mortara Garavelli, *Il discorso indiretto nell'italiano parlato*, in "Études Romanes", XXXIV (1995) 69-87.
1995b Bice Mortara Garavelli, *Il discorso riportato*, in *GGIC* III, Bologna, Il Mulino, pp. 426-468.
2001 Bice Mortara Garavelli, *Le parole e la giustizia,* Torino, Einaudi, 2001.
2003 Bice Mortara Garavelli, *Prontuario di punteggiatura*, Roma-Bari, Laterza, 2003.

MOTSCH - PASCH
1987 Wolfgang Motsch - Renate Pasch, *Illokutive Handlungen*, in "Studia Grammatica" XXV (1987) 11-79.

NØLKE
1994 Henning Nølke, *Linguistique modulaire: de la forme au sens*, Louvain - Paris, Peeters - Linguistique modulaire, 1994.

PERRET
1994 Michèle Perret, *L'énonciation en grammaire du texte*, ouvrage publié sous la direction de Claude Thomasset, Paris, Nathan, 1994.

PÉTILLON-BOUCHERON
2002 Sabine Pétillon-Boucheron, *Les détours de la langue. Étude sur la parenthèse et le tiret double*, Louvain - Paris - Dudley, Peeters, 2002.

RENZI - SALVI et alii
1988 *Grande grammatica italiana di consultazione*. Volume I, *La frase. I sintagmi nominale e preposizionale*, a cura di Lorenzo Renzi, Bologna, il Mulino, 1988.
1991 *Grande grammatica italiana di consultazione*. Volume II, *I sintagmi verbale, aggettivale, avverbiale. La subordinazione*, a cura di Lorenzo Renzi e Giampaolo Salvi. Bologna, il Mulino, 1991.

1995 *Grande grammatica italiana di consultazione*. Volume III, *Tipi di frase, deissi, formazione delle parole*, a cura di Lorenzo Renzi, Giampaolo Salvi e Anna Cardinaletti. Bologna, il Mulino, 1995.

CORPORA DI RIFERIMENTO.

Athenaeum Corpus	`http://www.bmanuel.org/projects/at-HOME.html`.
C-Oral Rom	`http://lablita.dit.unifi.it/coralrom/` (CRESTI - MONEGLIA 2005)
LISULB	Corpus privato del dipartimento di Linguistica Italiana dell'Università di Basilea. (`http://www.lisulb.unibas.ch/`)

Iørn Korzen
(Copenhagen Business School)

12. Mr. Bean e la linguistica testuale. *Considerazioni tipologico-comparative sulle lingue romanze e germaniche.**

0. PREMESSA. Verso la metà degli anni '90 un'equipe di linguisti dell'Università di Copenhagen e della Copenhagen Business School avviò un progetto di studio comparativo sulle lingue italiana e danese, più precisamente sulla produzione e strutturazione testuale nelle due lingue. Il punto di partenza della ricerca era un'ipotesi di differenze strutturali e di complessità sintattica e testuale, non solo a livello interlinguistico ma anche tra varianti diamesiche in entrambe le lingue. L'approccio metodologico era, da un lato, di carattere psicolinguistico-cognitivo: l'equipe si era ispirata teoricamente alla grammatica cognitiva di Langacker 1987, 1990 ed alla psicolinguistica testuale di Coirier - Gaonac'h - Passerault 1996, e come procedimento empirico si era progettata la creazione di una collezione di testi secondo il metodo dei *testi paralleli,* cioè testi autentici, prodotti in situazioni indipendenti ma simili nelle due comunità linguistiche e con un contenuto equivalente; tra le fonti di ispirazione vanno qui menzionati gli studi di Chafe 1980, di Tomlin 1987 e di Folman - Sarig 1990.

Invece per l'analisi dei dati i membri dell'equipe hanno seguito strade più o meno diverse ed indipendenti. Per me la base testuale è servita fra l'altro per la documentazione della relazione tra le dimensioni lessicale, morfologica e testuale che era stata ipotizzata nei lavori di un altro gruppo di linguisti, tutti della Copenhagen Business School, che nelle loro indagini comparative sulle lingue germaniche e romanze avevano seguito una pista lessicale-tipologica ispirata a studiosi come Talmy 1985, 2000 Vol. II. (per i verbi) e Pustejovsky 1995 (per i sostantivi). Cfr. il § 4 *infra*[1].

1. L'INDAGINE EMPIRICA: METODOLOGIA. Il modello psicolinguistico-cognitivo adottato dall'equipe responsabile per la creazione della collezione di testi assume come *tertium comparationis* il livello cognitivo, ovvero la *rappresentazione mentale* di input extralinguistici, e prevede due fasi o "dimensioni":

(1) la percezione: la fase che va da input extralinguistico a rappresentazione mentale;
(2) la testualizzazione: la fase che va da rappresentazione mentale a realizzazione e codificazione linguistica.

Per "input extralinguistico" si intende qualsiasi fatto, evento o circostanza cognitivamente registrabile e conservabile nel cervello sotto forma di rappresentazione mentale non-linguistica. Determinanti per la prima fase e per la rappresentazione mentale sono una serie di condizioni generali del locutore, fra cui le sue conoscenze enciclopediche e le capacità cognitive dipendenti da esse

*Ringrazio Elisa Corino e Marco Carmello per la ricerca di esempi nel corpus VINCA.

[1] Fra i corpora del gruppo di ricerca "L'italiano nella varietà dei testi", VINCA (Varietà di Italiano di Nativi Corpus Appaiato, reperibile all'indirizzo `www.corpora.unito.it`) è il più simile alla raccolta di testi studiata in questo contributo. Come "Mr. Bean" anche VINCA parte da un input iconico, una serie di vignette, ed anche VINCA è composto da testi narrativi di studenti universitari italiani italofoni di età compresa fra i 19 e i 25 anni. Essendo la raccolta di testi per VINCA iniziata nel 2005-2006, il gruppo si è concentrato sull'implementazione, sulle trascrizioni e sulla messa in rete dei materiali raccolti, operazione che finora ha toccato il 50% dei materiali. Poiché l'indagine linguistico-testuale degli aspetti trattati in questo contributo è in VINCA agli inizi, vengono qui forniti soltanto esempi di forme infinite del verbo e nominalizzazioni.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 209-224.

e dall'appartenenza a particolari tradizioni storiche e socioculturali. Tali capacità determinano le possibilità inferenziali (dette anche lo "schema cognitivo" o lo "script" di situazioni usuali e frequenti) nonché le possibili presupposizioni testuali e pragmatiche, come vedremo nel § 3.

Determinanti per la seconda fase, la testualizzazione, sono da una parte i tratti linguistici fissi dell'idioletto del locutore nonché, di nuovo, le sue conoscenze enciclopediche e abilità inferenziali, dall'altra i fattori specifici legati alla situazione comunicativa in questione. I fattori specifici includono l'interlocutore ed la relazione tra i comunicatori, il mezzo o canale della comunicazione, l'argomento e lo scopo del testo, il tipo e genere, ecc.

Il processo di testualizzazione prevede una fase strategica che varia quanto alla lunghezza e comprende la contestualizzazione, ovvero l'ancoraggio nel contesto referenziale e spazio-temporale, la scelta del macroatto linguistico (che può, evidentemente, variare ed essere più di uno nel testo complessivo) e la pianificazione testuale che include la scelta di variante diamesica (se essa non è già imposta dalla situazione), di variante diafasica, di forma, di contenuto, di lunghezza, e così via.[2]

2. LA CREAZIONE DELLA RACCOLTA DI TESTI. Essendosi basata sull'ipotesi della rappresentazione mentale come *tertium comparationis*, l'equipe aveva fatto il possibile perché i fattori relativi alla creazione dei testi italiani e danesi fossero i più simili possibili. I partecipanti erano 27 studenti italiani dell'Università di Torino e 18 studenti danesi dell'Università di Copenhagen, iscritti prevalentemente al primo od al secondo anno senza considerazioni di carattere sociale e, con pochissime eccezioni, dell'età tra i 19 ed i 25 anni.

Agli universitari, che non erano stati informati dello scopo dell'indagine, fu chiesto di raccontare il contenuto di due film brevi, più precisamente due episodi aventi per protagonista il personaggio di *Mr. Bean* impersonato dall'attore Rowan Atkinson: *The Library,* 'La biblioteca', della durata di 9 minuti, e *The Christmas Crib,* 'Il presepe' (che fa parte dell'episodio *Merry Christmas Mr. Bean* 'Buon Natale, signor Bean'), il quale dura 3 minuti. Le due sequenze sono molto diverse tra di loro, e come vedremo nel § 5, a volte la prima, a volte la seconda si presta meglio all'esame di uno specifico fenomeno linguistico.

Gli studenti furono divisi in due gruppi, così da avere materiale rappresentativo della lingua sia scritta che orale: il primo gruppo fu pregato di raccontare il primo episodio per iscritto ed il secondo oralmente e viceversa il secondo gruppo. I racconti orali furono registrati su nastro e successivamente trascritti, mentre per quelli scritti ogni partecipante aveva a disposizione un computer. Le istruzioni ai partecipanti erano le seguenti:

[1] GRUPPO A: Stai per vedere un video della durata di 3 minuti. È permesso prendere appunti. - Dopo che ne avrai preso visione, **racconta oralmente a una persona che non lo ha visto, ciò che è successo nel video**. Il tuo racconto verrà registrato su nastro.
(GRUPPO B - la stessa domanda per il video della durata di 9 minuti).
GRUPPO A: Vedrai adesso un altro video della durata di 9 minuti. Al suo termine dovrai **raccontare per iscritto ciò che è accaduto nel film**. NB Potrai prendere appunti durante la visione. Quando avrai terminato il tuo racconto, è permesso correggere lingua e/o contenuto, se lo riterrai necessario. (tempo: 45 minuti).
(GRUPPO B - la stessa domanda per il video della durata di 3 minuti). *Bean*, Istruzioni.

In questo modo fu creata una base testuale consistente di 54 testi italiani, 27 scritti e 27 orali, e di 36 testi danesi, 18 scritti e 18 orali. Si tratta di una collezione quantitativamente modesta, di circa 38.300 parole, ma qualitativamente interessante per la sua composizione di testi paralleli, italiani e danesi, scritti ed orali. Alla prima pubblicazione complessiva dell'equipe a cura di Skytte - Korzen - Polito - Strudsholm (di cui il terzo volume contiene l'intera collezione di testi

---

[2] Per più particolari si vedano Coirier - Gaonac'h - Passerault 1996, Skytte 1999, 2000, pp. 20 sgg., e Skytte - Korzen - Polito - Strudsholm 1999.

in versione cartacea), furono allegati i tre cd contenenti tutte le registrazioni orali, ed in seguito i testi scritti e trascritti sono stati messi a disposizione, in versione elettronica, di colleghi linguisti interessati.

Le motivazioni per scegliere le sequenze di *Mr. Bean* erano, da una parte, di carattere pratico: i prodotti filmici esistevano già e dai titolari dei diritti d'autore ottenemmo presto il permesso di usarli nel progetto. Dall'altra le sequenze erano molto adatte data la loro struttura: trattandosi di film praticamente muti, non si correva il rischio di "interferenza" linguistica. Inoltre, e di pari importanza, le sequenze non erano particolarmente marcate in senso culturale.

3. DIFFERENZE DI CONDIZIONI GENERALI E SPECIFICHE. Nonostante ciò abbiamo potuto osservare certi fenomeni che hanno avuto conseguenze per le testualizzazioni. Alcuni erano legati alle "condizioni generali" dei locutori, più precisamente alle loro conoscenze enciclopediche. Al momento della creazione dei testi la figura di *Mr. Bean* era generalmente più nota in Danimarca che in Italia. Ciò comportò, da parte dei partecipanti danesi, un maggiore grado di presupposizione di conoscenza, non solo della figura stessa ma anche del carattere umoristico delle sequenze. Invece in più casi, gli italiani ritennero necessario presentare più approfonditamente sia il protagonista che il genere artistico; cfr. l'esempio [14] sotto.

La stessa presupposizione di conoscenza si è potuta notare in alcuni dei resoconti orali: nonostante l'istruzione ai partecipanti di raccontare la storia del video oralmente "ad una persona che non lo ha visto", cfr. [1], sembra chiaro che in alcuni casi il parlante presupponga la conoscenza della figura da parte dell'interlocutore.

Viceversa si è potuta osservare una maggiore dimestichezza culturale della scena del presepe da parte degli italiani, i quali hanno fatto subito uso della parola *presepe* o *presepio.* Lo scenario del presepe è invece molto meno comune in Danimarca, e molti partecipanti danesi hanno adoperato descrizioni parafrastiche equivalenti a *rappresentazione della Natività, teatro dei burattini, esposizione di Natale, una specie di presepio,* ecc.

Nelle testualizzazioni della *Biblioteca* molti dei partecipanti italiani hanno ritenuto necessario spiegare il 'cigolare del pavimento' con il fatto che si tratta di un pavimento di legno. Invece ai danesi tale fatto è sembrato talmente evidente ed usuale da non meritare una menzione particolare (cfr. anche Skytte 1999).

Oltre a ciò abbiamo osservato certe differenze strategiche di testualizzazione dovute in parte alle condizioni generali dei locutori, in parte alle condizioni legate alla situazione comunicativa specifica. Si tratta di differenze di registro e di macroatto.

Generalmente il registro dei testi italiani, sia orali che scritti, è assai più alto di quello dei testi danesi. Sebbene in tutti e due si tratti di un'attività eseguita in università, non vi è dubbio che la scena universitaria comporti livelli di formalità diversi in Italia e in Danimarca. Inoltre molti degli studenti danesi si trovavano di fronte a docenti che già conoscevano, mentre gli italiani incontrarono un'equipe non solo di studiosi sconosciuti, ma di accademici stranieri. Non voler fare "brutta figura" ha indotto gli italiani in parte ad una variante diafasica più alta, in parte ad un macroatto diverso e legato anche alle condizioni generali menzionate poco sopra: data la minore notorietà di *Mr. Bean* in Italia, molti italiani hanno usato, almeno parzialmente, i macroatti *interpretare* ed *informare.* Invece i danesi, volendo fare anche loro "bella figura", hanno scelto più generalmente il *riferimento fedele* della trama, riferimento comprensivo di più elementi e dettagli possibili.

A tali differenze si aggiungono differenze generali legate alle tradizioni retorico-testuali delle due comunità linguistiche. Nel sistema scolastico italiano si dà più importanza alla "bella forma" di un testo, rispetto a quello che avviene in Danimarca, ed una "bella forma" richiede fra l'altro un alto grado di varietà sia stilistica che lessicale. Tratteremo le conseguenze linguistiche di queste differenze nel § 5.4.

4. TIPOLOGIA LINGUISTICA: LINGUE “ENDOCENTRICHE” E LINGUE “ESOCENTRICHE”. Malgrado le differenze menzionate nel paragrafo precedente e la dimensione modesta della collezione di testi, essa ha potuto servire convincentemente a documentare una serie di tendenze di strutturazione testuale causate dalle differenze tipologiche legate ai sistemi delle due lingue in questione, differenze descritte nei lavori dell’equipe della Copenhagen Business School.

Gli studiosi avevano constatato una diversa concentrazione semantica ed informativa nelle lingue germaniche e romanze: come tendenza generale i verbi delle lingue germaniche sono lessicalmente specifici e precisi, mentre i sostantivi sono relativamente generici. Data la concentrazione informativa nel verbo, ossia al centro della proposizione, tali lingue sono state denominate “endocentriche”. Invece nelle lingue romanze i sostantivi sono lessicalmente più specificati e precisi ed i verbi più generici; la concentrazione informativa è qui collocata negli argomenti nominali, ovvero al di fuori del centro della proposizione, motivo per cui queste lingue sono state denominate “esocentriche”. Fra le pubblicazioni del gruppo, cfr. soprattutto Korzen - Marello 2000, Herslund 2003, Baron 2003, Korzen - D’Achille 2005, Korzen 2004 e 2005a/b.

La specificazione lessicale dei verbi germanici dunque è dovuta alla lessicalizzazione, ovvero alla presenza nel lessema, della componente semantica MODO, vale a dire la maniera in cui si svolge l’azione verbale. Tale componente è invece generalmente assente nei verbi romanzi più frequenti. Buoni esempi sono qui i verbi di movimento: laddove per esempio un verbo tedesco o scandinavo con pochissime eccezioni non può fare a meno di esprimere il modo in cui si svolge il movimento,[3] cfr. esempi come *gehen, fahren, radeln / radfahren, segeln, reiten* – [danese] *gå, køre, cykle, sejle, ride,* tale componente semantica viene generalmente aggiunta al verbo romanzo sotto forma di satellite avverbiale: ‘andare *a piedi, in automobile, in bicicletta, in barca, a cavallo*’, e spesso non viene esplicitata. Insieme alla componente MODO questi verbi specificano anche la componente FIGURA, cioè il tipo di (s)oggetto coinvolto nell’azione verbale: i verbi citati richiedono tutti (s)oggetti (più o meno) particolari. Invece i verbi di movimento italiani più frequenti, *andare, venire, entrare, uscire, salire, scendere, partire, arrivare, tornare, cadere,* ecc. non specificano né MODO né FIGURA.

La specificazione lessicale dei sostantivi romanzi è dovuta alla loro tendenza a lessicalizzare la componente semantica FIGURA nel senso di forma o configurazione dell’oggetto in questione. Invece la componente semantica più frequentemente lessicalizzata nei sostantivi germanici è la FUNZIONE dell’oggetto.[4] Siccome oggetti che appaiono (più o meno) diversi fisicamente possono svolgere (più o meno) la stessa funzione, generalmente tale differenza comporta una lessicalizzazione più specifica, ovvero ad un livello iponimo, nelle lingue romanze rispetto alle lingue germaniche. In molti casi le lingue germaniche possono arrivare allo stesso livello di specificazione attraverso composizioni nominali, ma molto spesso appare soltanto la radice iperonimica e sottospecificata.[5]

In entrambi i ceppi linguistici la specificazione lessicale è dunque determinata dalla componente semantica FIGURA, ossia dall’*apparenza visuale* dell’azione o dell’entità in questione, e si può dire che le lingue romanze concepiscono e rappresentano il mondo extralinguistico come consistente di relazioni piuttosto generiche e astratte (denotate dai verbi) tra entità relativamente precise e specificate (denotate dai sostantivi), mentre le lingue germaniche concepiscono e rappresentano il mondo come consistente di relazioni piuttosto precise e specificate tra entità relativamente generiche e sottospecificate. Tali differenze tipologiche non si manifestano solo a li-

---

[3] A causa del lungo influsso del francese, l’inglese è piuttosto un misto di caratteristiche germaniche e romanze; cfr. Talmy 1985, 2001, Baron - Herslund 2005.

[4] La componente FIGURA nella terminologia dell’equipe danese corrisponde ai due “ruoli” (“qualia”) di Pustejovsky 1995, pp. 76-77, 85, il “constitutive role” e il “formal role”. Oltre a queste due componenti ed a FUNZIONE, Pustejovsky opera con l’“agentive role” dell’oggetto, che descrive *da chi* od *in che modo* l’oggetto è stato creato.

[5] Per più particolari sull’italiano e sul danese, cfr. Korzen 2004, 2005c.

vello lessicale, compaiono anche a livello sintattico e testuale: alla specificità lessicale dei verbi germanici tende a corrispondere anche une specificità grammaticale, cioè i verbi germanici tendono ad apparire in forme morfologiche che esplicitano più tratti grammatico-semantici possibili, vale a dire in forme finite. Viceversa alla genericità o sottospecificazione lessicale dei verbi romanzi tende a corrispondere una genericità e sottospecificazione anche grammaticale, cioè questi verbi tendono ad apparire molto più frequentemente in forme che non esplicitano tanti tratti grammatico-semantici, vale a dire in forme non finite o nominalizzate incorporate in un'altra proposizione matrice. La correlazione parallela si ritrova nei sostantivi: alla specificità lessicale dei sostantivi romanzi tende a corrispondere una specificità grammaticale, cioè questi sostantivi tendono ad apparire in sintagmi forniti di determinante, mentre molto più frequentemente i sostantivi sottospecificati germanici appaiono anche grammaticalmente sottospecificati, ossia in sintagmi senza determinante e (spesso) incorporati in una struttura verbale.[6]

Tali correlazioni sono in linea anche con le teorie di Hopper - Thompson 1980, 1984, che avevano constatato una diretta correlazione tra l'individuazione semantica di un costituente e la sua funzione testuale: più un costituente è semanticamente particolareggiato e visto come distinto dal suo background, e maggiore è la tendenza alla "funzione testuale prototipica" del costituente (Hopper - Thompson 1984, p. 708). La funzione testuale "prototipica" dei verbi è quella di istanziare indipendentemente[7] una "occorrenza" della classe, cioè un evento, un'attività od uno stato, funzione che richiede la forma verbale finita, mentre la funzione testuale prototipica dei sostantivi consiste nell'istanziare un'entità (del primo, del secondo o del terzo ordine nella terminologia di Lyons 1977, pp. 442 sgg.), il che richiede l'esplicitazione dei tratti espressi da un determinante. Hopper & Thompson operano monolinguisticamente, ma applicata comparativamente alle lingue endo- ed esocentriche la loro descrizione – verificata nelle tante lingue da loro analizzate – punterebbe ad una tendenza delle lingue endocentriche, fra cui il danese, alla "promozione" testuale dei loro costituenti verbali (più particolareggiati di quelli esocentrici) e, cioè, alla realizzazione in forma finita, mentre le lingue esocentriche, fra cui l'italiano, sarebbero tendenzialmente "programmate" a relegare i loro costituenti verbali ad un background testuale, eventualmente "incorporati" in un'altra struttura (frase matrice). Viceversa le lingue romanze sarebbero tendenzialmente programmate alla promozione testuale dei loro costituenti nominali (più particolareggiati di quelli endocentrici), e cioè alla realizzazione in sintagmi dotati di un determinante, mentre le lingue germaniche sarebbero tendenzialmente programmate a relegare i loro costituenti nominali ad un background testuale, eventualmente incorporati in un'altra struttura (struttura verbale).

Infatti, in linea di massima le differenze citate sono appunto tra le più notevoli dei due ceppi linguistici. E non solo: sembrano anche determinanti per certi sviluppi diacronici, illustrabili come nella tavola 1, *infra*

Più precisamente gli sviluppi diacronici sono i seguenti:

- nelle lingue scandinave, la perdita di molte forme verbali che nell'antico nordico esprimevano una subordinazione retorico-sintattica (il congiuntivo e molte forme e costrutti infiniti, cfr. anche il § 5.1), mentre molto più generalmente tali forme si sono mantenute nell'evoluzione dal latino alle lingue romanze, e
- nelle lingue romanze, il completamento relativamente veloce del sistema degli articoli a differenza delle lingue germaniche, dove mancano tuttora articoli indefiniti per i nomi massa ed al plurale.

[6] Per più particolari sul sistema nominale italiano e danese, cfr. Korzen 2005a/b.
[7] Cioè senza "appoggiarsi" ad un altro verbo come nel caso delle forme verbali infinite.

Specificità lessicale ↓		**Genericità lessicale** ↓	
Tendenza alla funzione "prototipica": all'istanziazione di una "occorrenza" della categoria in questione ↓		Tendenza alla funzione "atipica": alla non-istanziazione ed alla decategorizzazione ↓	
Specificità grammaticale		**Genericità grammaticale**	
Verbi germanici: tendenza alla finitezza	Sostantivi romanzi: tendenza alla determinazione	Verbi romanzi: tendenza all'infinitezza ed all'incorporazione	Sostantivi germanici: tendenza alla non-determinazione e/o all'incorporazione
Sviluppo diacronico			
Verbi germanici: sistema flessivo ridotto	Sostantivi romanzi: sistema di articoli completo	Verbi romanzi: ricchezza flessiva conservata	Sostantivi germanici: sistema di articoli incompleto

Tav. 1: lessico → testo → diacronia; specificità vs. genericità[8].

5. I DATI DI "MR BEAN". In questo paragrafo vediamo come i testi *Mr Bean* siano stati in grado di confermare le tendenze illustrate nella tavola 1, le quali possono essere così riassunte:

- nelle lingue romanze: una predisposizione allo stile nominale, alla deverbalizzazione,
- nelle lingue germaniche: una predisposizione allo stile verbale.[9] (Di nuovo va sottolineato che parliamo per sommi capi e di tendenze generali dei due ceppi linguistici, cfr. anche nota 8).

Di *deverbalizzazione* si può parlare in tutti i casi in cui una proposizione è stata realizzata senza verbo finito, vale a dire con un verbo infinito o nominalizzato oppure senza verbo, come una "frase ridotta". I casi di "frasi ridotte" (in senso un po' più lato del solito) possono essere suddivisi nei seguenti sottogruppi 2, 3 e 4, per cui conviene operare complessivamente con quattro tipi di deverbalizzazione:

1. casi in cui anziché un verbo finito appare un verbo infinito o nominalizzato, cfr. § 5.1;
2. predicativi liberi, cfr. 5.2;
3. apposizioni nominalizzate, cfr. 5.3;
4. anafore "infedeli" (con materiale lessicale diverso dall'antecedente), cfr. § 5.4.

5.1 FORME VERBALI INFINITE E NOMINALIZZATE. Vediamo prima i casi in cui anziché con un verbo finito, la proposizione è stata realizzata con un infinito, un gerundio, un participio od una nominalizzazione, come nei casi seguenti provenienti dall'insieme di testi che d'ora in poi per brevità sarà chiamato *Bean*[10]:

[2] mister Bean […] si è messo dei dei guanti, in modo *da non rovinare questo questo testo*
Bean, IMB9[11],

[8] Per più particolari, cfr. Korzen 2005b/c. È chiaro che vanno fatte le dovute riserve quanto alle differenze interlinguistiche all'interno di ogni ceppo linguistico nonché alle differenze intralinguistiche di carattere tipologico-testuale.
[9] Simmetricamente si può parlare di una *denominalizzazione* nelle lingue germaniche, la quale consiste in una riduzione retorico-testuale dei costituenti nominali. Tale riduzione è particolarmente evidente in danese, dove ha luogo nelle incorporazioni e in un tipo particolare di intransitivizzazione che ho descritto in Korzen 2005b.
[10] Per le sigle: *I* sta per italiano, *S* per testo scritto, *M* per testo orale; i partecipanti erano divisi in due gruppi, *A* e *B*. Nella trascrizione dei testi orali, la lineetta indica l'allungamento del suono precedente, e la virgola e i tre puntini indicano pause nel parlato rispettivamente brevi e lunghe.
[11] Un corrispettivo in VINCA può essere ravvisato nel seguente esempio:

[3] *Andando verso il tavolo dove può accomodarsi* l'ospite fa molta attenzione a non far scricchiolare il pavimento *Bean*, ISA1[12],

[4] *Arrivato al tavolo*, apre la sua borsa per prendere delle cose che a quanto pare gli serviranno. *Bean*, *ibid.*[13],

[5] Si accorge del*l'arrivo del bibliotecario,* quindi per non farsi scoprire in un certo senso, chiude il libro [...] *Bean*, IMB9[14].

Prima di tutto vanno fatte alcune premesse quanto alle differenze sistematiche, qui morfologiche, tra le nostre due lingue. In danese non esiste il gerundio né i costrutti participiali cosiddetti "assoluti" del tipo:

[6] *Vivente mia moglie,* spesso facevamo viaggi all'estero.[15]

[7] *Morto il re,* gli successe il figlio maggiore.[16]

Il costrutto in [6] è uno dei tipi scomparsi nell'evoluzione dall'antico nordico alle lingue scandinave attuali. Nella *Skånske lov*, 'la legge della Scania', che risale alla metà del Trecento, troviamo per esempio:

[8] *At bonda lifwande* ma aldrig kuna hans kæra af hans gerningum hwath sum han gør um henna egn. *Skånske lov*, ms. Stockholm B 69[17],

[8'] '*Vivente il marito,* sua moglie non deve mai lamentarsi delle sue azioni checché lui faccia degli averi di lei.'

[8''] *At* *bonda* *lifwande*
PREP marito.DAT vivente
'vivente il marito'
(in danese mod.: *Mens manden er i live,* letteralmente 'Mentre il marito è in vita'; cfr. n. 17).

[2b] `Dopo essere arrivato al pian terreno pianterreno percepì una strana sensazione ; come se qualcuno lo stesse seguendo ; questa diventò sempre più forte tanto` **`da farlo`** `spaventare così tanto che iniziò a correre , [...]` Vinca.

[12] Un corrispettivo in VINCA può essere ravvisato nel seguente esempio (si noti in questo caso la reduplicazione della costruzione):

[3b] **`Uscendo`** `dal bar dimentica il suo cappello sul tavolino . Non` **`accorgendosi`** `della sua dimenticanza comincia a camminare fino a quando non nota un' ombra di un uomo che lo sta inseguendo .` Vinca.

[13] Un corrispettivo in VINCA può essere ravvisato nel seguente esempio:

[4b] `Sorseggia amareggiato il suo boccale di birra e , appena` **`finito`** `, esce frettolosamente dalla locanda.` Vinca.

[14] È difficile trovare un corrispettivo per questo tipo di esempi, proponiamo come possibile esempio di questo fenomeno il seguente enunciato:

[5b] `non vedeva la sua famiglia da un mese e forse quasi sperava che al suo` **`arrivo`** `in stazione nessuno lo aspettasse.` Vinca.

[15] Analogamente in VINCA:

[6b] `il mio sguardo si fermò a fissare l' astuccio della mia scrivania` **`contenente`** `una penna azzurro cielo che avevo comprato in vacanza l' estate scorsa .` Vinca.

[16] Un costrutto temporale simile in VINCA potrebbe essere:

[7b] `appena` **`finito`** `, esce frettolosamente dalla locanda .` Vinca.

Si noti che la presenza dell'avverbio temporale rende più esplicita l'interpretazione in termini di *consecutio*.

[17] Il manoscritto "Stockholm B 69" è stato datato paleograficamente alla metà del Trecento; ma può trattarsi di una copia, per cui l'uso linguistico può essere più antico. In danese moderno avremmo:

[8'''] Mens manden er i live, må hans kone aldrig klage over hans gerninger, uanset hvad han gør med hendes ejendele.

Comunque, come si vede nella tavola seguente, che cita le percentuali medie calcolate su tutte le proposizioni della sequenza Bean *La biblioteca*, generalmente le forme infinite e le nominalizzazioni sono molto più rare nei testi danesi che in quelli italiani – ed in entrambe le lingue sono più rare nei testi orali che in quelli scritti[18]:

Proposizioni realizzate con:	Testi danesi		Testi italiani	
	orali	scritti	orali	scritti
Infiniti – cfr. es. [2]	6.40	12.02	20.10	23.98
Gerundi – cfr. es. [3]	–	–	6.37	14.39
Participi – cfr. es. [4]	0	0.01	0.62	5.77
Nominalizzazioni – cfr. es. [5]	0	0.01	0.10	2.97
Totale	**6.40**	**12.04**	**27.19**	**47.11**

Tav. 2: Proposizioni realizzate con verbo infinito o nominalizzate, *La biblioteca* (percentuali medie calcolate su tutte le proposizioni).

Andando da sinistra verso destra e cominciando dai testi orali danesi, pressappoco le occorrenze totali raddoppiano per ogni "salto" di tipo testuale, e nei testi scritti italiani arrivano a quasi metà delle proposizioni.

Come si è detto, i calcoli della tavola 2 sono basati sui resoconti della sequenza *La biblioteca,* la quale si adatta molto meglio ad una testualizzazione caratterizzata dai *rilievi testuali* prodotti dalle forme deverbalizzate. Invece l'altra sequenza, *Il presepe*, consiste di una serie di piccoli eventi narrativamente coordinati che non si prestano altrettanto facilmente alla distinzione tra primi piani e sfondi.

I rilievi testuali prodotti in questi casi sono causati dalla diversa specificità grammaticale delle forme coinvolte. Solo i verbi di "funzione prototipica" sono in grado di "istanziare" indipendentemente un evento verbale in un testo, cfr. il § 4. Invece alle forme deverbalizzate mancano i tratti grammatico-semantici tempo, modo, aspetto e soggetto,[19] i quali vanno interpretati con l'aiuto della frase matrice in cui è incorporata la proposizione deverbalizzata. In questo modo si crea il rilievo testuale in cui la proposizione realizzata con verbo finito in frase principale si trova posta in primo piano, esplicitando tutti i tratti necessari per l'interpretazione testuale, mentre le proposizioni senza verbo finito costituiscono lo sfondo.[20]

In questi casi, oltre alle differenze interlinguistiche, gioca un ruolo importante anche il registro, e, come accennato nel § 3, molti testi italiani sono caratterizzati da un registro alto e da una struttura piuttosto rigida e manchevole di dettagli. Un buon esempio è il seguente, un testo intero e quello più breve fra i testi italiani:

[9a] Nel filmato un noto comico inglese si reca nella sala di lettura di una biblioteca ***richiedendo*** un testo antico in visione. Nel silenzio assoluto che qui regna, inizia i suoi preparativi ***infastidendo*** il vicino, nonostante i suoi maldestri ***tentativi*** di evitare qualsiasi rumore. Proprio a causa delle ***occhiate*** torve dell'altro lettore, si distrae ***macchiando*** irrimediabilmente il libro. Ogni suo ***espediente*** per risolvere la situazione risulta controproducente.

---

[18] Alcuni esempi tratti da VINCA delle costruzioni discusse in questo paragrafo:

[9b] Disperato, ma sano e salvo, ritorna di corsa sui suoi passi **dirigendosi** alla sua abitazione — Vinca,

[9c] «Cosa c'è di meglio, in questa mattina, che passeggiare incautamente, **costeggiando** il treno fermo, odorante di legno marcio e ferro arrugginito» — Vinca.

[19] Tranne nei costrutti "assoluti" del tipo in [6]-[7], dove non manca il soggetto.

[20] Per più particolari, anche su altri tipi di rilievi testuali, cfr. anche Korzen 2002 e 2003.

> Decide, infine, di strappare le pagine ***rovinate***, ma ***rendendosi*** conto di ***averne staccate*** molte più del previsto, sostituisce il suo testo con quello del vicino momentaneamente ***distratto***. Riesce così a consegnare al bibliotecario un libro integro, ma rivela la sua colpevolezza ***tornando*** per recuperare il suo segnalibro personale ***dimenticato*** nell'opera ***danneggiata***. *Bean*, ISA6.

Come si vede, solo nove verbi (sottolineati) appaiono in forma finita, sette dei quali hanno *Mr. Bean* come soggetto; uno (*regna*) si trova in frase secondaria. Invece ben tredici verbi (in corsivo e grassetto) appaiono in forma infinita o nominalizzata[21]. I verbi finiti in frase principale costituiscono lo scheletro riassuntivo della storia a cui le forme deverbalizzate forniscono vari tipi di informazione suppletiva.

Il testo citato in [9a] consiste di 117 parole, mentre il testo scritto italiano più lungo, ISA4, consiste di 528 parole. Generalmente la stringatezza strutturale si rispecchia nella lunghezza del testo, ed i testi orali sono mediamente più lunghi di quelli scritti. La tavola seguente mostra le lunghezze medie dei quattro tipi di testo:

	Testi danesi		**Testi italiani**	
	orali	**scritti**	**orali**	**scritti**
Lunghezze medie (parole)[22]	1009,8	510,9	654,4	283,4

Tav. 3: Lunghezza media dei testi de *La biblioteca*.

In questo caso, andando da sinistra verso destra dai testi orali ai testi scritti di ogni lingua, i numeri non raddoppiano, bensì grosso modo si dimezzano. Sembra che si possa dire, quindi, che la frequenza di deverbalizzazione (Tav. 2) è più o meno inversamente proporzionale alla lunghezza del testo (Tav. 3)[23].

5.2 I PREDICATIVI LIBERI. I predicativi liberi (di solito del soggetto, più raramente dell'oggetto) esprimono una descrizione del soggetto (eventualmente dell'oggetto) legata e pertinente alla situazione designata dal verbo principale. Vi sono due tipi diversi, uno che – a differenza dei predicativi legati (argomentali) del soggetto e dell'oggetto – non fa parte della frase nucleare, la sua posizione è piuttosto libera, ed è parafrasabile con una frase gerundiva od avverbiale con valore temporale, causale, condizionale, modale, avversativo, o sim., con verbo copulativo ed il predicativo nella funzione di predicativo legato del soggetto. Se ne è già visto un esempio in [4] sopra, cfr. anche:

[10] *Cosciente del guaio creato* tenta l'ultima strada di salvezza *Bean*, ISA12,

[11] e così, *tutto spaventato* cerca- mm, cerca il modo- per, per ovviare a questo inconveniente *Bean*, IMB6[24].

Il secondo tipo fa parte della frase nucleare, la sua posizione è fissa dopo il verbo della frase ed il suo significato si avvicina a quello dell'avverbio; è comunque, anch'esso, parafrasabile con un costrutto con verbo copulativo:

[12] […] gli viene consegnato il volume richiesto. Indossati i guanti bianchi, messo il segnalibro, comincia a sfogliarlo *felice*. [~ ed è felice] *Bean*, ISA14,

---

[21] Fra le nominalizzazioni ho incluso forme come *occhiate (torve)* e *(ogni suo) espediente,* equivalenti a costrutti finiti come *(L'altro lettore) lo guarda (torvamente)* e *Tutto quello che fa.*

[22] Sono consapevole di tutte le riserve che bisogna fare con un calcolo in base all'unità "parola"; però in qualunque modo si fosse eseguito il calcolo, la differenza proporzionale risultava quella illustrata.

[23] Invece nei racconti del *Presepe* non c'è la stessa differenza a causa della struttura narrativa diversa della sequenza. Le lunghezze medie sono queste: testi danesi orali – scritti: 283,8 – 296,6; testi italiani orali – scritti: 217,0 – 255,2. La maggiore lunghezza dei testi scritti in questi casi è in larga misura dovuta alle migliori condizioni mnemoniche che hanno portato all'inclusione di più dettagli narrativi.

[24] Circa l'uso delle lineette e delle virgole, cfr. nota 10.

[13] strappa le pagine, sempre cercando di nascondersi dal suo vicino che lo guarda un po' insospettito [~ ed è un po' insospettito][25] *Bean*, IMB4.

Appunto per il costituente verbale mancante questi costituenti contribuiscono allo stile nominale ed alla stringatezza testuale. Per la loro natura non deve quindi sorprendere che troviamo esattamente le stesse tendenze di distribuzione che abbiamo visto nel § 5.1 (anche se il primo tipo, di [10]-[11], si manifesta molto più frequentemente del secondo). In questo caso ho calcolato la frequenza media dei costituenti per mille parole nei quattro tipi di testo:

	Testi danesi		**Testi italiani**	
	orali	**scritti**	**orali**	**scritti**
Predicativi liberi	0,22‰	2,39‰	1,41‰	5,80‰

Tav. 4: Numero di predicativi liberi per mille parole nei testi de *La biblioteca*.

5.3 LE APPOSIZIONI NOMINALIZZATE. Anche fra le apposizioni ci sono costrutti nominalizzati. L'apposizione si distingue dal predicativo libero per trovarsi sempre posposta al costituente di cui esprime una descrizione od elaborazione, e la descrizione od elaborazione espressa non è limitata alla situazione designata dal verbo della frase. L'apposizione costituisce sempre un sintagma o frase a sé stante, ed in Korzen, in stampa, ho proposto una descrizione secondo cui può trattarsi di frasi principali a verbo finito, di frasi subordinate a verbo finito, di frasi a verbo infinito e di sintagmi nominali, aggettivali o preposizionali. Per il nostro contesto sono particolarmente pertinenti le frasi a verbo infinito ed i sintagmi menzionati, come per esempio:

[14] Durante il filmato abbiamo assistito alle azioni compiute da un individuo che […] si diverte a giocare, modificando il classico scenario del presepe natalizio. Questo individuo, *noto comico televisivo e esponente dello humour anglosassone,* manovra e anima le statuine, dando loro vita e voce, […] *Bean*, ISB7,

[15] Un primo piano inquadra subito un presepe, *costruito in maniera assai tradizionale:* vi è la capanna con la paglia e la culla […] *Bean*, ISB12.

Come già svelano i due esempi citati, in questi casi ho scelto di fare l'analisi sulla sequenza *Il presepe*. Dato che gli elementi descrivibili ne *La biblioteca* sono pochi, 4-5 per essere precisi: Mr. Bean, l'altro lettore, il bibliotecario ed i due libri consultati, le apposizioni occorrenti in questi testi sono altrettanto poche ed un calcolo della media avrebbe dato un'immagine molto insicura. Invece le apposizioni nei testi del *Presepe*, anche se sempre di un numero piuttosto modesto, permettevano il calcolo statistico con un margine più elevato di sicurezza.

In un primo momento sembrava che le occorrenze appositive dovessero contraddire l'immagine vista fino a questo punto (il numero fra parentesi è il numero totale di occorrenze):

	Testi danesi		**Testi italiani**	
	orali	**scritti**	**orali**	**scritti**
Apposizioni	13,75‰ (35)	13,49‰ (36)	19,74‰ (60)	23,21‰ (77)

Tav. 5: Numero di apposizioni per mille parole e totali nei testi del *Presepe*.

[25] Ho descritto il secondo tipo, con il termine "predicativo secondario", come un caso di "valenza derivata" in Korzen 1996, pp. 215-216. Regula - Jernej 1975, pp. 298-300, chiamano entrambi i tipi "predicativi liberi", mentre *GGIC* II, pp. 196 e sgg., 208 e sgg., distingue tra "frasi ridotte" nel primo caso e "complementi predicativi del soggetto accessori" nel secondo.

Ma una veloce analisi della tipologia appositiva rivela presto un particolare interessante, anche se non sorprendente: la maggior parte delle apposizioni – sia italiane che danesi – sono, infatti, frasi relative, del tipo:

[16] Sulla scena del presepe passa prima una banda su un carro, poi compare un gregge di pecore, *che viene portato via a bordo di un autocarro.* *Bean*, ISB8.

Il tipo in [16], frequente appunto nei testi narrativi, non ci interessa in questo contesto perché non dice nulla a proposito di predisposizioni o meno allo stile nominale. La tavola 6 dimostra i casi di frasi relative appositive,

	Testi danesi		**Testi italiani**	
	orali	**scritti**	**orali**	**scritti**
Frasi relative appositive	12,92‰ (33)	9,74‰ (26)	16,46‰ (50)	16,58‰ (55)

Tav. 6: Numero di frasi relative appositive per mille parole e totali nel *Presepe*.

e la tavola 7 i casi rimanenti, tutti di apposizioni senza verbo finito, ovvero apposizioni nominalizzate; infatti, le frasi relative costituiscono l'unico tipo di apposizione a verbo finito nella collezione di testi:

	Testi danesi		**Testi italiani**	
	orali	**scritti**	**orali**	**scritti**
Apposizioni nominalizzate	0,78‰ (2)	3,75‰ (10)	3,29‰ (10)	6,63‰ (22)

Tav. 7: Numero di apposizioni nominalizzate per mille parole e totali nel *Presepe*.

Va detto che le occorrenze sono molto poche, ma l'immagine generale di una maggiore tendenza alla nominalizzazione in italiano che in danese, ed in testi scritti che in quelli orali, si mantiene inalterata.

5.4 ANAFORE "INFEDELI". L'ultimo elemento che contribuisce alla stringatezza testuale ed alla ricchezza informativa sono le anafore "infedeli", le anafore lessicalmente diverse dai loro antecedenti.[26] La scelta tra anafore "fedeli" (lessicalmente identiche) ed infedeli dipende sia dalla tipologia testuale che dalle norme e tradizioni retoriche. La variazione lessicale è particolarmente frequente in testi narrativi, giornalistici e saggistici e meno frequente in testi tecnici e giuridici in cui vige il principio di univocità e precisione. Interlinguisticamente la variazione lessicale è, *ceteris paribus*, molto più frequente in italiano, dove i diversi cambiamenti stilistici sono generalmente desiderati ed apprezzati, che in danese, dove vige piuttosto il concetto del "parlar chiaro", un concetto che comprende fra l'altro la ripresa di un referente con lo stesso materiale lessicale dell'antecedente. Però pure in danese si osserva un uso non infrequente di anafore infedeli in testi giornalistici e saggistici.

In tutti i casi in cui sia antecedente che anafora sono sintagmi nominali (antecedente SN_1, anafora SN_2) viene presupposta – ma non esplicitata – la predicazione:

[17] Il SN_1 è un N_2.

Senza l'accettazione di tale predicazione, la catena anaforica non "funziona" testualmente. La predicazione può essere a priori garantita per motivi lessicali, ciò avviene nei casi $N_1 = N_2$ (i due

[26] Il termine "anafora infedele" appartiene alla tradizione francese ed è stato suggerito da Blanche-Benveniste - Chervel 1966, pp. 30-31.

sono identici, cfr. [18a], o sinonimi, cfr. [18b]), $N_1 < N_2$ (N_2 è un iperonimo di N_1, cfr. [18c]) e $N_1 > N_2$ (N_2 è un iponimo di N_1, cfr. [19]):

[18] Ho visto *un'automobile* nel nostro cortile ieri sera. […] [a] *L'automobile* / [b] *La macchina* / [c] *Il veicolo* era di una marca che non conosco.

[19] Ho visto *un veicolo* nel nostro cortile ieri sera. C'era qualcosa di strano in *quell'automobile.*[27]

La predicazione in [17] può essere assicurata anche per motivi pragmatici, più precisamente per conoscenze enciclopediche, cfr. [20], per conoscenze oggettivamente verificabili di carattere *ad hoc* (condivise o meno dall'interlocutore), cfr. [21a], o per una valutazione soggettiva da parte del parlante, cfr. [21b]:

[20] Stasera arriva Umberto Eco. *Lo scrittore italiano* si ferma fino a domenica.

[21] Stasera arriva Luca Orsi. [a] *Il mio compagno di scuola* / [b] *Il mascalzone* si ferma fino a domenica.

Lo stesso SN anaforico può contenere informazione sia oggettiva che soggettiva:

[22] Stasera arriva Umberto Eco. *Il brillante scrittore italiano* si ferma fino a domenica.

Nei racconti de *La biblioteca* le anafore fedeli ed infedeli si distribuiscono percentualmente in questo modo:

	Testi danesi		**Testi italiani**	
	orali	**scritti**	**orali**	**scritti**
Anafore fedeli	94,2	90,4	81,8	59,5
Anafore infedeli	5,8	9,6	18,2	40,5

Tav. 8: Tipologia anaforica nei testi de *La biblioteca*, percentuali medie.

È interessante notare il quasi raddoppiamento delle anafore infedeli andando, come prima, da sinistra verso destra nella tavola. Il tipo più frequente dipende dall'antecedente: in tutti i casi sono usuali le anafore oggettivamente informative, cfr. [21a], nel caso di Mr. Bean però (non sorprendentemente) miste con le valutazioni soggettive (le quali appaiono anche in un paio di casi sull'altro lettore), cfr. [21b], e nel caso del libro miste con anafore (quasi) sinonimiche:

a. su Mr. Bean: *l'uomo, il nostro protagonista, il nostro, il personaggio, il tipo che entra, questo personaggio dall'aspetto molto buffo, il comico personaggio, il buffo personaggio, il pazzerello, il poverino;*

b. sull'altro lettore: *il (suo) vicino, il vicino di banco/di posto, il signore (che gli sta) accanto, l'uomo seduto di fronte, l'altro ospite della biblioteca, il lettore, l'altro lettore, l'altro studioso, l'altra persona, l'ignaro signore, il severo signore, il poverino;*

c. sul bibliotecario: *l'addetto alla biblioteca, la persona addetta, il responsabile della biblioteca, il signore che controlla la biblioteca, il guardiano (della biblioteca), l'assistente, l'inserviente, il custode;*

d. sul libro / sui libri: *il libro, il testo, il testo rovinato, il testo completamente distrutto, l'opera, il volume, il manoscritto, le due pergamene, il tomo, questo manuale, il prezioso volume, il libro preziosissimo;*

L'anafora enciclopedica, cfr. [20], appare solo in un paio dei testi danesi dove anziché *Mr. Bean* troviamo *Rowan Atkinson* nel riferimento alla stessa figura.

[27] Sulle condizioni particolari delle riprese iponimiche, cfr. Korzen 2001, 2006.

In tutti i casi l'anafora infedele risparmia l'esplicitazione della predicazione in [17] e funziona in questo modo, si può dire, come elemento deverbalizzante.

6. CONCLUSIONE. Direi che i testi di *Mr. Bean* hanno più che chiaramente confermato l'ipotesi iniziale, cioè che l'italiano, in quanto lingua esocentrica, ha chiare predisposizioni verso uno stile nominale, mentre il danese, lingua endocentrica, *ceteris paribus,* è caratterizzato da uno stile più verbale. È importante sottolineare che causa delle diverse predisposizioni è solo in parte la tipologia linguistica, fondamentale è inoltre il macroatto impiegato nella testualizzazione. Come già detto, molti degli italiani hanno ritenuto necessario interpretare, spiegare e informare sull'input extralinguistico, necessità non sentita altrettanto fortemente dai danesi, che per questo hanno potuto concentrarsi sul semplice riferire, narrare la storia, macroatto che non solo "permette", ma prescrive lo stile verbale.

Posso comunque aggiungere che tendenzialmente altri corpora (vedi gli esempi di Vinca in nota) confermano l'immagine delineata in questo intervento.

Inoltre, ispirati dall'indagine *Bean,* molti miei laureandi hanno fatto simili ricerche su altri tipi e generi di testo, fra cui testi giornalistici, testi giuridici, newsgroup e siti web, ed in tutti i casi l'immagine generale è stata confermata. Più alto è il registro (come per esempio nei testi giuridici), e maggiore è la differenza tra i testi italiani e danesi; più basso è il registro (come nei newsgroup), e più i testi italiani e danesi si somigliano.

BIBLIOGRAFIA.

BARON

2003 *Language and Culture*, edited by Irène Baron, Copenhagen, Samfundslitteratur, 2003 "Copenhagen Studies in Language" 29.

BARON - HERSLUND

2005 Irène Baron - Michael Herslund, *Languages endocentriques et langues exocentriques. Approche typologique du danois, du français et de l'anglais*, in HERSLUND - BARON 2005, pp. 35-53.

BLANCHE-BENVENISTE - CHERVEL

1966 Claire Blanche-Benveniste - André Chervel, *Recherches sur le syntagme substantif*, in "Cahiers de lexicologie" IX (1966)[2] 3-37.

CHAFE

1980 *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, editor Wallace L. Chafe, Norwood (N.J.), Ablex Pub. Corp., 1980.

COIRIER - GAONAC'H - PASSERAULT

1996 Pierre Coirier - Daniel Gaonac'h - Jean-Michel Passerault, *Psycholinguistique textuelle. Approche cognitive de la compréhension et de la production des textes*, Paris, Armand Colin, 1996.

D'ACHILLE

2004 *Generi, architetture e forme testuali. Atti del 7. Convegno SILFI, Società internazionale di linguistica e filologia Italiana (Roma, 1-5 ottobre 2002)*, a cura di Paolo D'Achille, Firenze, Franco Cesati, 2004 "Università Roma Tre, Dipartimento di italianistica; SILFI, Società internazionale di linguistica e filologia italiana".

EGERLAND - WIBERG
2003 *Atti del 6. Congresso degli italianisti scandinavi. Lund, 16-18 agosto 2001*, a cura di Verner Egerland e Eva Wiberg, Lund, Romaniska Institutionen - Lunds Universitet, 2003, pp. 313-324.

FOLMAN - SARIG
1990 Shoshana Folman - Gissi Sarig, *Intercultural Rhetorical Differences in Meaning Construction*, in "Communication and Cognition" XXIII (1990)[1] 45-92.

GGIC I → RENZI - SALVI et alii 1988; II → RENZI - SALVI et alii 1991; III → RENZI - SALVI et alii 1995.

HERSLUND
2003 *Aspects linguistiques de la traduction*, sous la dir. de Michael Herslund, Pessac, Presses Universitaires de Bordeaux, 2003. [Textes des communications présentées lors d'un colloque franco-danois organisé à l'Université Michel de Montaigne-Bordeaux III, 30-31 mai 2001, par l'Équipe de recherche en syntaxe et sémantique de l'Université Michel de Montaigne-Bordeaux III et l'Équipe de recherche Traduction et linguistique de la Faculté de langues de la Haute École d'études commerciales de Copenhague, Université Michel de Montaigne-Bordeaux III, 30-31 mai 2001].

HERSLUND - BARON
2005 *Le génie de la langue française. Perspectives typologiques et contrastives*, édité par Michael Herslund et Irène Baron, Paris, Larousse, 2005 "Langue française" 145.

HOPPER - THOMPSON
1980 Paul J. Hopper - Sandra A. Thompson, *Transitivity in Grammar and Discorse*, in "Language" LVI (1980)[2] 251-299.
1984 Paul J. Hopper - Sandra A. Thompson, *The Discourse Basis for Lexical Categories in Universal Grammar*, in "Language" LX (1984)[4] 703-752.

JANSEN et alii
2002 *L'infinito & oltre. Omaggio a Gunver Skytte*, a cura di Hanne Jansen, Paola Polito, Lene Schøsler e Erling Strudsholm, Odense, Odense University Press, 2002.

KORZEN
1996 Iørn Korzen, *L'articolo italiano tra concetto ed entità. Uno studio semantico-sintattico sugli articoli e sui sintagmi nominali italiani con e senza determinante – con un'indagine particolare sulla distribuzione del cosiddetto "articolo partitivo"*. Vol. I. *Considerazioni preliminari. Il sintagma nominale senza determinante*, Vol. II. *I sintagmi nominali con articolo. Conclusioni*, København, Museum Tusculanum Press, 1996 "Études Romanes" 36.
2001 Iørn Korzen, *Anafore e relazioni anaforiche. Un approccio pragmatico-cognitivo*, in "Lingua nostra" LXII (2001)[3-4] 107-126.
2002 Iørn Korzen, *Il trapassato prossimo in un'ottica pragmatico-testuale*, in JANSEN et alii 2002, pp. 203-226.
2003 Iørn Korzen, *Rilievi testuali nel sistema verbale: un panorama comparativo*, in EGERLAND - WIBERG 2003, pp. 313-324.
2004 Iørn Korzen, *Dalla microstruttura alla macrostruttura*, in D'ACHILLE 2004, pp. 363-376.
2005 *Lingua, cultura e intercultura: l'italiano e le altre lingue, atti del VIII Convegno SILFI, Società Internazionale di Linguistica e Filologia Italiana (Copenaghen, 22-26 giugno 2004)*, a cura di Iørn Korzen, Copenaghen, Samfundslitteratur Press, 2005 "Copenhagen Studies in Language" 31.

2005a Iørn Korzen, *Struttura linguistica e schema cognitivo: tipologie a confronto*, in KORZEN 2005, pp. 123-134.

2005b Iørn Korzen, *Lingue endocentriche e lingue esocentriche: lessico, testo e pensiero*, in KORZEN - D'ACHILLE 2005, pp. 31-54.

2005c Iørn Korzen, *Linguistic Typology in Translation: Endocentric and Exocentric Languages, as Exemplified by Danish and Italian*, in "Perspectives. Studies in Translatology" XIII (2005)[1] 21-37.

2006 Iørn Korzen, *On Demonstrative Determiners in Anaphoric Noun Phrases*, in NØLKE et alii 2006, pp. 261-277.

i.s. Iørn Korzen, *L'apposizione, un costituente trascurato*, in "Studi di Grammatica Italiana" XXIV (2005), in corso di stampa.

KORZEN - D'ACHILLE

2005 *Tipologia linguistica e società. Considerazioni inter- e intralinguistiche. Due giornate italo-canadesi di studi linguistici: Roma, 27-28 novembre 2003*, a cura di Iørn Korzen e Paolo D'Achille, Firenze, Franco Cesati, [2005] "Copenhagen Business School - Università Roma Tre, Dipartimento di italianistica".

KORZEN - MARELLO

2000 *Argomenti per una linguistica della traduzione | Notes pour une linguistique de la traduction | On Linguistic Aspects of Translation*, a cura di Iørn Korzen e Carla Marello, Alessandria, Edizioni dell'Orso, 2000 "Gli argomenti umani" 4.

LANGACKER

1987 Ronald W. Langacker, *Foundations of Cognitive Grammar.* Vol. I. *Theoretical prerequisites*, Stanford, Stanford University Press, 1987.

1991 Ronald W. Langacker, *Concept, Image and Symbol. The Cognitive Basis of Grammar*, Berlin - New York, Mouton de Gruyter, 1990.

LYONS

1977 John Lyons, *Semantics*, Voll. 1-2, Cambridge - London - New York - Melbourne, Cambridge University Press, 1977.

NØLKE et alii

2006 *Grammatica. Festschrift in honour of Michael Herslund | Hommage à Michael Herslund*, eds. | éds. Henning Nølke, Irène Baron, Iørn Korzen, Hanne Korzen and | et Henrik Høeg Müller, Bern, Peter Lang, 2006.

PUSTEJOVSKY

1995 James Pustejovsky, *The Generative Lexicon*, Cambridge (Massachusetts) - London (England), The MIT Press, 1995.

REGULA - JERNEJ

1975 Moritz Regula - Josip Jernej, *Grammatica italiana descrittiva su basi storiche e psicologiche*, Bern - München, Franke Verlag, 1975_2 [1965_1].

RENZI - SALVI et alii

1988 *Grande grammatica italiana di consultazione*. Volume I, *La frase. I sintagmi nominale e preposizionale*, a cura di Lorenzo Renzi, Bologna, il Mulino, 1988.

1991 *Grande grammatica italiana di consultazione*. Volume II, *I sintagmi verbale, aggettivale, avverbiale. La* subordinazione, a cura di Lorenzo Renzi e Giampaolo Salvi. Bologna, il Mulino, 1991.

1995 *Grande grammatica italiana di consultazione*. Volume III, *Tipi di frase, deissi, formazione delle parole*, a cura di Lorenzo Renzi, Giampaolo Salvi e Anna Cardinaletti. Bologna, il Mulino, 1995.

SHOPEN
1985 *Language Typology and Syntactic Description*, Vol. 1. *Clause Structure*, Vol. 2. *Complex Constructions*, Vol. 3. *Grammatical Categories and the Lexicon*, edited by Timothy Shopen, Cambridge (Cambridgeshire) - New York, Cambridge University Press, 1985.

SKYTTE
1999 Gunver Skytte, *Il progetto "Mr. Bean in danese e in italiano"*, in SKYTTE et alii 1999, pp. 10-33.
2000 Gunver Skytte, *Sprogbrug i komparativt perspektiv,* in SKYTTE - KORZEN 2000, pp. 13-64.

SKYTTE et alii
1999 *Strutturazione testuale in italiano e in danese. Risultati di una indagine comparativa | Tekststrukturering på italiensk og dansk. Resultater af en komparativ undersøgelse*, a cura di | redigeret af Gunver Skytte, Iørn Korzen, Paola Polito & Erling Strudsholm, Copenhagen | København, Museum Tusculanum Press, 1999.

SKYTTE - KORZEN
2000 Gunver Skytte - Iørn Korzen, *Italiensk-dansk sprogbrug i komparativt perspektiv: reference, konnexion og diskursmarkering*, Frederiksberg, Samfundslitteratur, 2000.

TALMY
1985 Leonard Talmy, *Lexicalization patterns: semantic structure in lexical form*, in Shopen 1985, Vol. 3, pp. 57-149.
2000 Leonard Talmy, *Toward a Cognitive Semantics*, Vol. 1. *Concept Structuring Systems*, Vol. 2. *Typology and Process in Concept Structuring*, Cambridge (Massachusetts), MIT Press, 2000 [2001[r]].

TOMLIN
1987 Russell S. Tomlin, *Linguistic Reflections of Cognitive Events*, in TOMLIN 1987a, pp. 455-479.
1987a *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984*, edited by Russell S. Tomlin, Amsterdam - Philadelphia, John Benjamins, 1987.

CORPORA E SITI DI RIFERIMENTO.

Mr. Bean `http://frontpage.cbs.dk/MrBean-korpus/`

VINCA `http://www.bmanuel.org/projects/vn-HOME.html.`

ELISA CORINO
(Università di Torino)

13. NUNC est disputandum.
Aspetti della testualità e questioni metodologiche.

POL. *Though this be madness,*
yet there is method in't.
William Shakespeare, *Hamlet*, II.2.

0. PREMESSA. La lingua in rete pone al linguista difficoltà a livello sia teorico sia metodologico perché comporta l'approccio a fenomeni in costante evoluzione e spesso difficilmente riconducibili ai paradigmi della linguistica tradizionale. È il caso, in particolare, dei *newsgroup*, che costituiscono la base della suite di corpora NUNC (Newsgroup UseNet Corpora) e che sembrano sfuggire a definizioni di "testo" rigidamente ancorate a quelle canoniche, elaborate per *Textsorten* tradizionali.

Il contributo si propone di fornire una prima soluzione ai nodi problematici sollevati dai NUNC e di mostrare alcune delle possibili applicazioni di ricerca più innovative. Così, dopo aver presentato le peculiarità dei newsgroup e dei NUNC, se ne fa seguire un'analisi da una prospettiva precipuamente testualista, quella che ci è parsa più adatta a coglierne la specificità tra i tipi di testo della Comunicazione Mediata dalla Rete (CMR).

I corpora NUNC meritano attenzione anche perché offrono la possibilità di condurre analisi lessicografiche e statistiche all'interno di materiale linguistico di recentissima datazione: nella seconda parte, pertanto, si affrontano le fondamentali questioni metodologiche, proponendo un'esemplificazione (basata su una recente tesi di laurea: Casavecchia 2005) per lo studio delle collocazioni nella terminologia specialistica.

0.1 I NEWSGROUP QUESTI SCONOSCIUTI: CHI SONO, COME FUNZIONANO. Quando tra gli anni '70 ed '80 del secolo scorso nacque in America la rete UseNet, nessuno avrebbe immaginato l'enorme successo, la rapida diffusione e la portata dell'impatto sull'evoluzione della lingua in tutte le sue sfaccettature che tale mezzo di comunicazione avrebbe avuto negli anni a venire.

L'idea originaria alla base della creazione di UseNet era quella di dare vita ad una rete che mettesse in contatto gli utenti di Unix e che servisse da punto di riferimento per chiunque avesse domande o problemi ad esso connessi. Nelle sue prime forme essa si poneva come «a poor man's ARPANET» (Hauben 1997), un'alternativa gratuita all'elitaria rete ufficiale. Fin da sùbito dunque si instaurarono le condizioni che ancora oggi contraddistinguono i newsgroup[1] (Newsletter Group o gruppi di discussione, d'ora in poi, per brevità, NG) in quanto forma peculiare della Comunicazione Mediata dalla Rete (CMR)[2]: una comunità - virtuale - di persone che

[1] Si noti che tanto in questo articolo come nel resto del presente volume si predilige il prestito inglese alla sua traduzione italiana. Questo in parte per ragioni storiche: si tratta di una forma di comunicazione nata in ambiente anglofono e giunta a noi solo nei tardi anni '90; in parte per consuetudine degli autori, abituali fruitori della rete e familiari con le sue convenzioni linguistiche; in parte ancora per rispettare lo scioglimento dell'acronimo NUNC (Newsgroup UseNet Corpora).

[2] Condividiamo qui il punto di vista di Allora 2005 che (sulla scorta di Herring 1996 e Rheingold et alii 1994) individua entro una generale CMC (Comunicazione Mediata dal Computer) un àmbito di comunicazione più ristretto, la CMR (Comunicazione Mediata dalla Rete), che raccoglie e-mail, Internet Relay Chat, NewsGroup, Multi User Dungeon, blog, mailing list, forum, ecc., ma che ignora i testi statici come, ad esempio, i siti web.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 225-252.

condividono gli stessi problemi ed interessi e che si creano un proprio spazio, estraneo ad altri canali, per poterne discutere. UseNet è stata spesso comparata ad una serie di riviste specializzate che, dall'impulso primigenio dato dagli argomenti suggeriti dagli utenti Unix, si sono rapidamente diversificate per genere e soggetti discussi fino ad arrivare a comprendere discussioni di filosofia, cucina, scienze… In un certo senso il contribuire alla creazione ed allo sviluppo di un newsgroup dà ad ognuno l'opportunità di avere un proprio spazio di visibilità, una sorta di casa editrice privata presso la quale "pubblicare" pensieri e discussioni. Molte delle caratteristiche dei newsgroup hanno fatto sì che questi venissero assimilati spesso a dei tazebao virtuali, a delle bacheche telematiche nelle quali affiggere messaggi[3]. A ben guardare, però, le possibilità offerte da un newsgroup vanno ben oltre il semplice scambio di informazioni: in primo luogo perché lo scambio di informazioni avviene all'interno della bacheca stessa secondo la modalità uno-a-tanti (il messaggio non è indirizzato ad un individuo né ad un elenco postale, ma all'argomento di dibattito e può essere letto da tutti coloro che condividono lo spazio virtuale, o, per dirla nel gergo del caso, che *postano* sul newsgroup); in secondo luogo perché il forum di discussione è articolato in una tassonomia precisa, «ossia in un sistema di cornici argomentative che si chiamano "gerarchie", a base geografico-nazionale e/o tematica che, peraltro, nascono dal basso in base alla iniziativa degli utenti» (cfr. Barbera ¶ 1, § 2.2.5, in questo volume).

(1)	comp.*	*Computer topics, both hardware and software.*
(2)	news.*	*Administration of the Big 8, as well as about Usenet and Netnews in general, and related topics.*
(3)	sci.*	*Science and technology.*
(4)	humanities.*	*The humanities.*
(5)	rec.*	*Recreational topics, including music, sports, games, outdoor recreation, hobbies, crafts, ...*
(6)	soc.*	*Socializing, society, and social issues.*
(7)	talk.*	*Endless discussion, largely about politics.*
(8)	misc.*	*A mixture of newsgroups that don't fit the other 7 hierarchies. Many are about the practical aspects of everyday life.*

Tav. 1. Le *Big 8*
(da `http://www.big-8.org/dokuwiki/doku.php?id=history:big-8`).

I nomi dei newsgroup di UseNet definiscono una gerarchia, con il punto, ".", usato come separatore tra i suoi differenti livelli, come accade anche per i nomi di dominio. A differenza però di quanto avviene per questi ultimi, qui la parte più significativa del nome è messa per prima. Questa parte è dunque speciale e più significativa rispetto al resto, dal momento che definisce il più alto livello della gerarchia UseNet a cui quel gruppo appartiene. Per quanto riguarda le gerarchie tematiche si identificano quelle che tradizionalmente vengono chiamate le *Big8 Hierarchies*[4] (cfr. Tav. 1, *supra*). A base geografico/nazionale invece i nomi delle gerarchie iniziano

[3] Questo, naturalmente, è solo uno degli usi possibili di un newsgroup, forse il più tipico; ma molto può variare da gerarchia a gerarchia, giungendo anche a gruppi il cui uso medio è assai più prossimo ad una chat (ad es. `bln.jugend.talk`, `free.it.4amicialbar`, ecc.).

[4] Le *Big8 Hierarchies* (è diffusa anche la grafia *heirarchies*, in cui complice alla metatesi sembra essere l'incrocio "popolare" con *heir* 'erede" e derivati) sono il frutto di una ristrutturazione di UseNet avvenuta nel 1987, comunemente conosciuta con il nome di *Great Renaming*. La principale ragione della riorganizzazione fu la difficoltà di tenere sotto controllo e gestire il numero sempre crescente di newsgroup che proliferava in rete

con il codice ISO del paese ospitante, abbiamo così `it.diritto`, `it.diritto.condominio`, `it.diritto.assicurazioni`, ...; `it.discussioni.animali`, `it.discussioni.animali.gatti`, `it.discussioni.animali.cani`, ..., `it.discussioni.auto`, `it.discussioni.auto.ford`, ...; ecc.

Pur non essendoci un "gestore di Usenet", vi sono procedure e consuetudini, che la comunità si è data per mantenere le gerarchie, che vanno sotto il nome di *Netiquette*; in particolare esistono regole precise per le RFD (*Request For Discussion*) e le CFV (*Call For Votes*), ovvero i passi formali per "proporre" un nuovo gruppo Usenet; ogni NG ha poi di norma ha un manifesto (*charter*) che aiuta i nuovi arrivati (*newbies*) a comprendere quali sono gli argomenti oggetto di discussione e come trattarli. La conversazione procede attraverso l'invio di *articoli* o *post*[5] – strutturati come catene di post in sequenza (*thread*), ordinate in base al loro titolo (*subject*)[6].

1. NEWSGROUP, UN NUOVO CONCETTO DI TESTO? L'oggetto di questa discussione ruota intorno alle caratteristiche di un particolare tipo di comunicazione: i newsgroups, un dominio testuale ancora poco studiato rispetto ai suoi più celebri "cugini" - la chat, le e-mail ed il "contenitore" ipertesto; eppure creatività e salienza informativa rendono i NG un campo di indagine linguistico degno di grande attenzione[7].

Negli ultimi anni la linguistica testuale si è spesso interrogata sull'opportunità di rinnovare il concetto di testo alla luce delle nuove opportunità offerte dai media, che hanno imposto una riflessione sulla validità delle definizioni tradizionali. Il ruolo giocato dal medium è strettamente legato alle caratteristiche mutanti della norma e del sistema linguistico usato: il computer permette la commistione di elementi a tal punto che la principale e prototipica caratteristica dei testi prodotti nella CMC è proprio l'ibridità. Non si tratta qui esclusivamente della possibilità di creare ipertesti ed oggetti semiotici *à la* Petőfi: la questione, infatti, investe anche la dicotomia diamesica tra scritto e parlato.

(resoconti più dettagliati delle ragioni del *Great Renaming* si trovano nei post di Gene Spafford a `net.news` e `net.news.group` reperibili sul sito `http://groups.google.com/groups?selm=4558%40gatech.CSNET` e tra le varie FAQ sul *Great Renaming* online alla pagina `http://www.linux.it/~md/usenet/gr.html`). I newsgroup vennero così categorizzati in sette grandi gruppi tematici (`comp.*`, `misc.*`, `news.*`, `rec.*`, `sci.*`, `soc.*`, e `talk.*`) ai quali a metà degli anni '90, in seguito all'enorme espansione della rete UseNet, venne aggiunta `humanities.*`.

Alle *Big8 Hierarchies* si aggiunge poi un'altra gerarchia che, diversamente dalle prime otto, non è soggetta a procedure di controllo e organizzazione: la gerarchia `alt.*` (si tratta dell'abbreviazione di *alternative*, ma spesso è considerata sinonimo di anarchia: «The name *alt* was said to refer humorously to "anarchists, lunatics, and terrorists", but is understood by most people today as an abbreviation of "alternative"» Wikipedia s.v. *alt.* hierarchy*), sorta di buco nero senza regole in cui raggruppare tutto ciò che esula dalle gerarchie regolate. Nonostante alcuni evidenti aspetti negativi, `alt.*` offre anche notevoli vantaggi, tra i quali quello di ospitare newsgroup di argomenti molto specifici che non troverebbero altrimenti altra collocazione.

Del problema dei gruppi "binari", infine, tacciamo, dato che non ci riguardano in questa sede. Basti dire che di regola l'invio di materiali non testuali (file, immagini, ecc.) è limitato a gruppi (`*.binaries.*`, appunto) appositi, facilitando in ciò il compito anche di chi, come noi, è interessato al solo materiale testuale.

[5] Così vengono chiamati i messaggi in questo contesto per differenziarli da quelli di posta elettronica, *[e-]mail*.

[6] È possibile accedere ai gruppi tramite un portale web come Google, Arianna ed Usenetportal, o più direttamente con un programma (*newsreader*) dedicato, come ad esempio Agent (o Free Agent) della Forté.

[7] Ci risulta ad oggi un solo corpus – nell'accezione indicata da Barbera - Corino - Onesti ¶ 3 –predecessore dei NUNC: ELWIS, *Korpusgestütze Entwicklung lexikalischer Wissensbesen*, corpus creato presso l'università di Tübingen nel 1993, che raccoglieva un'annata delle gerarchie tedesche `cl.*`, `cnet.*`, `de *`, `fido.*`, `maus.*`, `stgt.*` e `zer.*`, nel complesso 647 NG per un totale di 43.300 articoli (43 milioni di parole, 540.000 types): cfr. Hinrichs et alii 1995 e Feldweg - Kibiger - Thielen 1995. L'unico altro precedente, per quel che ci è noto, è il *CMU Text Learning Group Data Archive* noto come "20 Newsgroups", una collezione di 20.000 post scaricati nel 1993 da 20 newsgroup organizzata da Tom Mitchell come base per *machine learning* (cfr. Mitchell 1997); che però, stando ai criteri di cui sopra, non può intendersi come un corpus.

La CMC permette, rispetto alla comunicazione scritta tradizionale, la riduzione degli intervalli temporali di comunicazione; rispetto alla comunicazione orale, tuttavia, i tempi di azione-reazione sono più dilatati. Scritto e parlato, registro formale ed informale, lontananza e vicinanza, pianificazione ed immediatezza, distanza e legami sociali, sono caratteristiche che definiscono due poli estremi di un continuo, all'interno del quale si individuano i diversi generi testuali – se di generi testuali si può parlare – inscritti nella rivoluzione nata dall'uso dei nuovi media. CMC è inoltre un concetto estremamente generale che racchiude al suo interno tanti (troppi?) modi di comunicazione, dalla chat all'e-mail passando per gli spazi virtuali di MUD (Multi User Dimension) e MOO (Multi-user-dimension Object Oriented), che difficilmente possono essere considerati espressione di un unico "genere testuale" in virtù di caratteristiche comuni (eccezion fatta per il medium che utilizzano ed alcuni elementi di intertestualità ricorrenti).

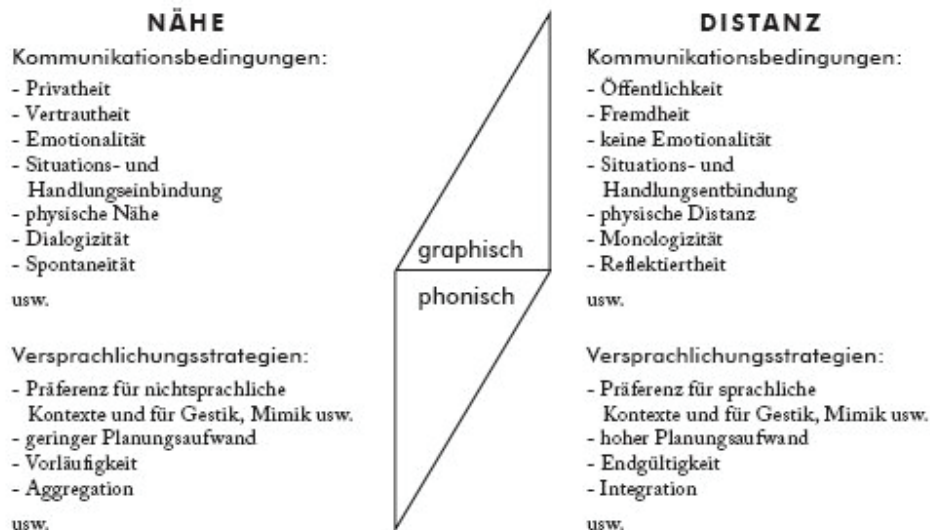
È d'altra parte vero che testi "multimediali" che fanno ricorso alla commistione dei codici esistono da sempre, così come all'interno dei vari generi possiamo distinguere registri diversi a seconda della situazione in cui il testo si inscrive.

Messo a confronto con le tipologie di comunicazione che fanno uso della rete, un newsgroup è mero testo che non gode di altre caratteristiche del Web: comunicazione visiva nulla; poca multimedialità (almeno fuori dai gruppi binari); basso livello di interattività con la macchina; presenza solo occasionale di link ipertestuali[8]. Anche la velocità di comunicazione tanto determinante nelle chatline e certo discriminante della CMC in genere, non conta molto nell'ambiente che andremo ad analizzare, in cui in genere le risposte, date offline e poi postate, si "prendono il loro tempo" e non sono legate al "perdere il turno" come in chat.

Mancando queste caratteristiche, dicevamo, la proprietà preponderante resta il testo, un tipo di testo che si sviluppa su più interventi di più interlocutori, mostrando un processo di testualizzazione peculiare rispetto ad altri *Textsorten*, come ben si vedrà nei §§ sgg.

1.1 NEWSGROUP, TRA SCRITTO E ORALE. Haase et alii 1997 considerano i newsgroup come forme di comunicazione essenzialmente scritte, mentre altri (Storrer 2000, Crystal 2001) sono concordi nel ricondurli ad una posizione molto vicina al parlato spontaneo. Senza dubbio entrambe le posizioni forniscono una buona analisi della natura dei newsgroup, anche se parziali. In quanto forme di comunicazione asincrona, i cui tempi possono spesso avvicinarsi a quelli della comunicazione sincrona, che si svolge in un certo ambiente e che presuppone l'interazione di più parlanti, il testo dei newsgroup si avvicina molto alle caratteristiche tipiche dell'oralità, eppure, come nota Storrer 2000, si tratta pur sempre di un testo scritto che non di rado assume le qualità specifiche della comunicazione scritta. Questo comporta alcune significative conseguenze, come fa d'altra parte notare Feenberg 1989, p. 23 (cfr. Lenke - Schmitz 1995, p. 121), «For example, we may no longer assume that writing is more formal and less personal than speech. This and other strange consequences must be taken into account in any online setting».

Anche Fiorentino 2004, riprendendo il modello di Koch - Österreicher 1994, sottolinea il carattere ibrido di alcuni mezzi di CMR: dal punto di vista del medium vengono realizzati con codice grafico, dal punto di vista concettuale si assiste all'oscillazione tra tratti di immediatezza e di distanza. L'immediatezza comunicativa è legittimata in particolare dai casi di comunicazione sincrona nelle chat e - aggiungiamo noi - nei newsgroup ad alto indice dialogico-comunicativo; la distanza è però chiaramente presente per la "non compresenza fisica degli interlocutori", per l'uso di un software che non consente che si realizzi un feedback simultaneo nelle interazioni o che si rispetti l'adiacenza dei turni".

[8] È comunque pur vero che riferimenti ad altro materiale online si fanno più frequenti in newsgroup specialistici, in cui gli utenti fanno uso frequente di collegamenti ipertestuali: nei newsgroup di fotografia si commentano foto presenti su questo o quell'altro sito, in quelli di motori si inseriscono i rimandi per poter vedere i dettagli delle componenti meccaniche, in altri ancora si inseriscono le URL da cui scaricare programmi e altri materiali …

Tav. 2. Il modello di Koch - Österreicher 1994, p. 588.

Ci sembra tuttavia che il riferimento alla lingua orale abbia spinto i linguisti ad un'attenzione spesso fuorviante ed anomala, ad analisi troppo legate alla spontaneità (vera o presunta) del discorso, alla presenza di elementi quali interiezioni, ideofoni, espressioni gergali o volgari, emoticons (certamente tentativi di rendere alcuni tratti del discorso orale, cosa che ha per l'italiano valenza particolare, visto lo sviluppo diacronico diversificato che lingua orale e scritta hanno seguito), ma che colgono solo una dimensione stilistico-espressiva superficiale, forse importante ma non esaustiva. Cfr. infatti la nozione di *Umgangssprache* (cfr. ad esempio Spitzer 1922/2007) cui ricorre Barbera ¶ 1, § 2.2.5.

Tav. 2. Il modello tratto da Lenke - Schmitz 1995, p. 120.

Come si può notare dalla tavola soprariportata, UseNet è il solo medium che insieme alla IRC consente una comunicazione *m:m* ("molti a molti"); si distingue dalla comunicazione orale perché è scritta, dalla comunicazione orale e dall'IRC in quanto è asincrona.

La variazione di registro e l'oscillazione tra maggiore e minore formalità nei NG dipende da due fattori principali.

Da una parte il senso di comunità che i partecipanti condividono, ovvero quella rete di comunicazione online organizzata ed autodefinita per interesse o scopo comune, che ha sede nella piazza virtuale. La percezione di appartenere ad un gruppo ben definito (familiarità ed interesse comune) induce una "libertà di movimento ed espressione" che favorisce la scomparsa di perifrasi, introduzioni all'argomento o lunghe spiegazioni (il tema si intuisce facilmente dal titolo dato al thread ed è facile inserirsi in una conversazione di cui sono stati "registrati" tutti i passaggi) e contemporaneamente la comparsa di modalità di interazione, formule di saluto e commiato, routine, che diventano caratteristiche di quel particolare gruppo e contribuiscono a dare la cifra delle relazioni all'interno della comunità (e di individuare chi a questa comunità è estraneo, come troll e niubbi).

D'altro canto la gerarchia e gli argomenti che nel newsgroup vengono discussi esercitano un peso decisivo nel determinare la varietà di lingua utilizzata. Come campioni rappresentativi di altrettante possibili varietà presenti nella rete, in questo lavoro si sono scelte cinque gerarchie di newsgroup: `it.comp.grafica.photoshop` (cfr. ess. [1, 8-11]), `it.cultura.storia.moderato` (cfr. es. [2]), `free.it.4amicialbar` (cfr. ess. [3-5]), `it.arti.musica.classica.mod` (cfr. es. [6] e Tav. 8) ed `it.arti.scrivere` (cfr. es. [7] e Tav. 7).

Come campioni di lingua controllata ed adeguata a tematiche specialistiche o complesse, abbiamo selezionato due NG di fotografia digitale e storia (`it.comp.grafica.photoshop` [1], `it.cultura.storia.moderato` [2]), dei quali il primo è tra i più tecnici ed è quindi esemplificativo della casistica più settorialmente marcata:

[1]
```
Potresti scoprirlo solo dal valore cromatico dei pixel
interessati dall' " ombra " , che nel caso del drop shadow
standard sono affetti da un " multiply " , mentre nella mia
ipotesi si ha solo una sovrapposizione di livelli senza
interazione . Credo sia uno studio molto ardito ... Mi è venuto
istintivamente da optare per il glow , perché non è molto
intuitivo assegnare una distanza pari a 0 a un' ombra .
```
it.comp.grafica.photoshop (NUNC-IT Photo).

[2]
```
Il papa Innocenzo III ( Lotario Conti , 22 febbraio 1198 - 16
luglio 1216 ) nel IV Concilio Laterano , tenuto dall' 11 al 30
novembre 1215 , si scagliò contro la corruzione con queste
parole : [...]
```
it.cultura.storia.moderato (NUNC-IT Generic I).

Il caso opposto è rappresentato dagli ess. [3-5], tratti da `free.it.4amicialbar`, forum di intrattenimento molto simile ad una chat, ricco di abbreviazioni, emoticons, interiezioni, acronimi, alta dialogicità e velocità nel botta e risposta.

[3]
```
ops era un re ... cosa cera prima che ti ruzzoli dal ridere ?
:-)
```
free.it.4amicialbar (NUNC-IT Generic II).

[4]
```
>>> credo ci fosse l' orzo ! (^_^)
>> bene vurria mai ... ;-)
> ma bere si , ti capita ! :)
in questo bar no ! è una vergogna chi non serve chi si fa fuori
tutto già prima di offrirlo ... ma va bene così per i
trigligeridi ;-)
```
free.it.4amicialbar (NUNC-IT Generic II).

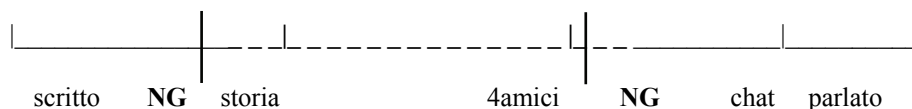
[5]
```
> ma non sei già impegnata ? O_O
si ma mi impegno molto :-P
poi sai siamo come le leonesse ;-)
```
free.it.4amicialbar (NUNC-IT Generic II).

Tra le caratteristiche di questa tipologia di conversazione ritroviamo amplificata al massimo (quasi a livello di chat) una caratteristica che è propria di numerosissimi newsgroup: il *topic shift*, laddove in NG come `free.it.4amicialbar` non c'è in realtà un vero e proprio argomento di cui parlare. Il fenomeno si osserva anche in altri thread e gerarchie, ma si resta generalmente all'interno del macroargomento del newsgroup ed in genere il nuovo topic si ricollega in modo logico a quello precedente.

Ad un livello intermedio tra questi due estremi, i NG `it.arti.musica.classica.moderato` ed `it.arti.scrivere` costituiscono caratteristici esempi di *scrittura dell'uso medio* (Baracco 2002), mostrando al loro interno un'ulteriore differenziazione che ha reso adeguata una classificazione in cinque ipotetici gruppi:

[6]
```
Per un musicista Sibelius va benissimo . Su questo siamo d'
accordo. Ma per un grafico ( e per l' editoria in generale )
Finale è di gran lunga meglio perché ti lascia maggiori libertà
. Per quanto riguarda l' aspetto grafico delle pagine fatte con
Sibelius , devo dire che a me non piacciono granché : le note
sono troppo " patatose " . Ma qui si scende nel personale . :-)
> Sugli errori sintattici di cui parli non comprendo proprio a
cosa ti riferisca . Cosa vuol dire che Finale non permette di
fare errori sintattici ? In che senso ?
```
it.arti.musica.classica.mod (NUNC-IT Generic I).

[7]
```
Il fan di SW come tale deve accettare tutte le boiate che Lucas
partorisce dalla sua mente , per ridurre il tutto alla sagra
dell' effetto speciale fine a se stesso ?
Uaz , e pensavo foste persone serie ... :]P
```
it.arti.scrivere (NUNC-IT Generic I).

Nell'opposizione oralità-scrittura possiamo identificare un complesso di situazioni comunicative che partecipano in gradi diversi all'oralità ed alla scritturalità "prototipiche", secondo una struttura a parentesi simile allo sviluppo delle scatole cinesi: etichette generali contengono elementi diversi che, a loro volta, si articolano in ulteriori sottocategorizzazioni. All'interno della categoria newsgroup, ad esempio, identifichiamo una sorta di scala che va da una maggiore formalità (vicinanza allo scritto) ad una maggiore oralità (dialogicità del parlato).

Tav. 3. Gradienti "scalari" scritto-parlato nei newsgroup.

Riguardo alla dicotomia scritto-parlato, Berruto 1985, p. 146, osservava che l'italiano parlato avrebbe la stessa grammatica dello scritto, soltanto più liberalizzata e più focalizzata sul parlante (e meno sul sistema); il parlante reinterpreterebbe le regole grazie alla presenza di un contesto chiarificatore: c'è, infatti, una vicinanza spaziale – virtuale nel nostro caso – che normalmente manca in una situazione di lingua scritta. I newsgroup presentano quella stessa liberalizzazione della lingua ed orientamento verso l'utente del parlato, ma in un ambiente scritto, dove viene meno una reale vicinanza spaziale, ma si ritrova quella virtuale di cui sopra, che certamente accorcia le distanze e sposta il sistema *ego-hic-nunc*, ma che pure manca delle circostanze extralinguistiche in cui i parlanti sono abitualmente immersi nella loro "normale" comunicazione orale.

Se in una conversazione orale possiamo individuare fenomeni di ripresa, ripetizione e ridondanza che aiutano gli interlocutori al superamento della mancanza di un testo a cui riferirsi e fungono da indicatori nel processo di focalizzazione, la natura scritta del newsgroup permette di mantenere costantemente esplicito l'insieme di entità ed oggetti che costituiscono il centro dell'attenzione dei parlanti/scriventi attraverso l'uso del quoting: i messaggi di un newsgroup presentano cioè una costante ripresa citazionale del testo originale di messaggi precedenti o di parti di essi, di solito visibilmente riconoscibili perché accompagnati dal segno di maggiore ">"[9] all'inizio di ogni riga di testo riportata.

Il meccanismo della focalizzazione, inoltre, cambia a seconda che si tratti di casi di dialogo *task oriented* o di conversazione casuale. Nei primi si presume che tutte le enunciazioni siano rilevanti nel senso che tutto il loro contenuto proposizionale veicoli informazione necessaria. Nella conversazione casuale invece non esiste l'obbligo per un ascoltatore di orientare la propria attenzione sullo stesso elemento focalizzato dal parlante. Si tratta infatti di un evento linguistico principalmente collaborativo in cui l'interesse per un determinato elemento da parte di un parlante non basta a far sì che questo diventi rilevante all'interno del dialogo: l'interesse deve essere negoziato da tutti i partecipanti al dialogo, affinché venga condiviso. Il dialogo all'interno dei newsgroup condivide entrambi i casi: thread incentrati su un tema particolare che richiedono risposte puntuali (una particolare funzione di Photoshop, il percorso per raggiungere un ristorante…) mostrano tutte le caratteristiche del dialogo *task oriented*, altri thread (non necessariamente appartenenti a newsgroup di "conversazione") presentano invece conversazione casuale, ed altri ancora alternano le due. Anche laddove le conversazioni appaiono più spontanee e "rilassate" nella forma, i newsgroup sono comunque maggiormente strutturati in termini testuali, sia per la gerarchizzazione del dialogo sia per il fatto che la conversazione avviene "in differita": si può quindi parlare di una testualità ragionata.

Soprattutto però ci sembra che l'autore di un intervento passi spesso in secondo piano rispetto al contenuto dell'intervento stesso. In un'ottica di analisi delle caratteristiche scritto-parlato è interessante notare come i partecipanti si accostino al newsgroup con intenzioni comunicative che non corrispondono (come avviene invece nella chat) alla volontà di riprodurre un dialogo faccia-a-faccia: per la funzione stessa che è alla base della nascita di UseNet gli utenti si alternano spesso "senza volto", prestando attenzione soprattutto a *ciò* che è scritto e non sempre avvertendo come importante *chi* l'ha scritto, se non in casi specifici in cui gli utenti "veterani" o più attivi si scambiano battute anche di tipo personale – sempre però in aggiunta, o di seguito, al commento dell'argomento del post, o dopo aver dato l'informazione richiesta dall'altro partecipante: non c'è mai spazio per comunicazioni puramente personali nel newsgroup "medio"[10].

Arno Scholz 2003 fa inoltre notare che la volatilità e l'instabilità dei generi testuali elettronici sono condizioni che predispongono ad una certa noncuranza verso le norme della lingua scritta. Se le caratteristiche della scrittura digitale favoriscono una certa libertà, non bisogna dimenticare che le scelte stilistiche e normative dipendono di gran lunga dallo scrivente e molto meno dal mezzo, ed il controllo dello scritto è forse proprio un metodo per veicolare un certo desiderio di ufficialità. Ciò non toglie che solitamente la precisione delle osservazioni tecniche conviva con un linguaggio assai sciolto, colloquiale, ricco di interiezioni o volgarismi.

[9] Che, anzi, è addirittura organizzato in serie, a volte abbastanza lunghe, a seconda del grado di quoting. La caratteristica è molto regolare in quanto frutto meccanico di una impostazione che il newsreader applica automaticamente ad ogni *reply* che l'utente compie; certo, come tutte le impostazioni è modificabile, e non mancano utenti che impostano ("settano") un diverso default: ma per fortuna sono pochi.

[10] E, naturalmente, diciamo *medio* perché una certa cautela, legata all'eterogeneità dei tipi di testo compresenti nei newsgroup (cfr. *supra*), è d'obbligo: i gruppi d'intrattenimento come `free.it.amici`, `bln.jugend.talk`, `free.it.4amicialbar`, ecc., sono molto più sbilanciati sulla comunicazione personale.

[8a]
```
> Ho tentato solo di essere più preciso , non prenderla come una
provocazione .
eddai avevo messo pure la faccina divertita ... ¬___¬
Il drop shadow continua , a mio giudizio , ad essere più
malleabile ...
distanza 0 ? embè ? feather e risolvi . Non basta ? Gaussian
Blur e passa la paura ... ;)
```
it.comp.grafica.photoshop (NUNC-IT Photo).

[8b]
```
Ma guarda quei banner li ho messi così a cazzo tanto per dargli
una parvenza figa ...... potevo anche non metterli ... lol ...
Ciao
```
it.comp.grafica.photoshop (NUNC-IT Photo).

1.2 NEWSGROUP E MASSIME CONVERSAZIONALI. Analizzando alcuni thread, notiamo inoltre come la comunicazione tipica dei NG proceda generalmente in conformità delle classiche massime conversazionali proposte da Grice 1967, che, infatti, permettono la comunicazione mirando insieme all'efficacia ed all'efficienza comunicativa.

1	QUANTITÀ	fornisci un contributo tanto informativo quanto richiesto
2	QUALITÀ	dì ciò che ritieni essere vero
3	RELAZIONE	sii pertinente
4	MODO	sii perspicuo/efficace.

Tav. 4. Le massime conversazionali di Grice 1989, p. 62.

Nei NG, queste massime si possono riconfigurare, riformulare e chiosare al modo seguente:

(1) L'informazione è massimizzata: si risponde ad una specifica domanda generalmente senza dare più informazioni del dovuto (gli interventi possono essere anche di poche righe, sul Web si preferisce una comunicazione veloce); se invece ciò succede è per favorire la massima (4).

(2) Si danno generalmente informazioni vere; se si presentano notizie poi smentite da altri, emerge spesso in un secondo tempo che l'informazione errata non era intenzionale. Non dimentichiamo che molti sono esperti del proprio settore, e chi ne sa di meno è di solito colui che fa domande od aggiunge commenti soggettivi, laddove le risposte più tecniche sono lasciate a chi ritiene di poter replicare in maniera attendibile.

(3) Pertinenza delle domande all'argomento del newsgroup.
Pertinenza delle risposte ai messaggi precedenti: questa è favorita dalla ripresa (integrale o parziale) dei testi altrui – la replica segue quasi sempre il frammento di testo a cui essa si collega – anche se alcuni scriventi non risparmiano divagazioni, in particolare per accattivarsi la simpatia della lista o per sollevare ulteriori questioni.

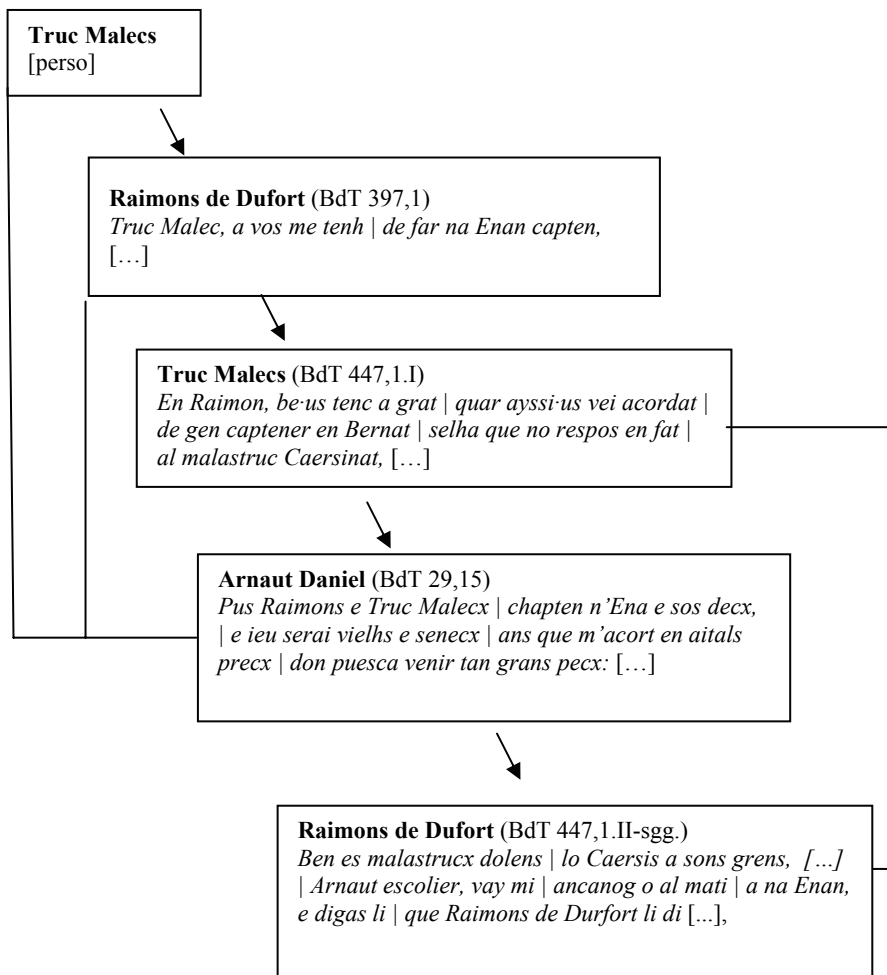
(4) Efficacia: sempre con le dovute riserve legate ai singoli utenti od argomenti ed ai numerosi *topic shift*, ci sembra che di solito lo scambio proceda in modo mirato alla soddisfazione delle richieste. Chi può fornire dati in più lo fa anche a qualche giorno di distanza (senza problemi di comprensione perché si riporta il testo di riferimento) ed in maniera chiara: da manuale in alcuni casi, col vantaggio però di essere un manuale *user-friendly* poiché costruito con un linguaggio "ordinario".

Un efficace esempio di quanto detto in (4) viene dal gruppo di Photoshop, ed è diviso per chiari punti anche se la terminologia è affatto tecnica:

[9]
```
1) selezione rettangolare sull' area scelta
2) menù contestuale --> layer via copy
```

```
3) Layer style : drop shadow , sfuma con 15 di raggio
4) seleziona il layer inferiore e aggiungi un Fill Layer con
viraggio verso i toni del verde. Se necessario riduci l' opacità
```

it.comp.grafica.photoshop (NUNC-IT Photo).

1.3 IDENTIFICAZIONE DEL TESTO E COERENZA. Il testo seguente, un celebre e “famigerato” scambio di sirventesi del XII secolo[11], mostra come la struttura dialogica attribuita unicamente a certa parte dei “nuovi testi” (e-mail, chat, newsgroup) non sia poi così “nuova” e si ritrovi anzi in un genere ben più antico. In questo testo, infatti, non solo gli autori intrattengono una discussione a distanza su un tema preciso ed il testo cresce e si alimenta della loro interazione fatta di risposte a interventi precedenti, ma è ben evidente anche la multiautorialità nel suo complesso[12].

Tav. 5. La tenzone tra Truc Malecs, Raimons de Durfort ed Arnaut Daniel su Na Ena (testi da Contini 1936, pp. 228-30 ed Eusebi 1984, p. 4).

[11] Ringrazio Manuel Barbera per aver richiamato la mia attenzione su questi testi.
[12] E non è un caso che la tradizione manoscritta abbia avuto difficoltà a mantenere la separazione tra i vari testi ed i vari autori, completamente restaurata solo in sede filologica (cfr. Contini 1936).

Una risposta scritta da due interlocutori in parallelo: a secoli di distanza dall'età contemporanea si ritrova già un modello di testualità che si contraddistingue nelle linee di intreccio e sviluppo della storia, secondo un filo conduttore che va oltre il singolo partecipante. Gli interventi vengono da più parlanti; condividono uno stesso tema, ovvero l'opportunità o meno di accettare la ipotetica "prova d'amore" richiesta da Na Ena; fanno riferimento all'argomento con riprese lessicali (*malastruc*, *malastrucx*) e di rima, o con varie allocuzioni agli interlocutori – *"Truc Malec, a vos me tenh ..."*; *"En Raimon ..."*, ecc. – (cfr. Contini 1936).

Nei newsgroup lo sviluppo sequenziale degli scambi è tradotto (come si è tentato in Tav. 5 di mostrare anche per i sirventesi occitanici) in modo visivo, grazie all'uso di appositi newsreader, programmi applicativi per la lettura dei thread, per cui è possibile identificare lo svolgersi del discorso nella sua cronologia e nel suo naturale avvicendamento di autori. Agent (e Free Agent) della Forté, ad esempio, organizza i messaggi per soggetti e visualizza all'interno di ogni thread l'ordine ed il reciproco riferimento dei post.

Lines	Subject (All Messages)	Author	Date ▲
8	**Arial Unicode**	Nikki	06/11/2004 0.10
10	Alan J. Flavell		06/11/2004 0.25
15	Nikki		09/11/2004 1.11
14	Alan		06/11/2004 16.45
21	Nikki		09/11/2004 1.09
15	Alan		09/11/2004 12.46
21	Nikki		10/11/2004 1.05
24	RSD99		11/11/2004 0.14
11	Andreas Höfeld		11/11/2004 10.47
20	RSD99		11/11/2004 20.57
23	Alan		12/11/2004 9.50
6	RSD99		12/11/2004 18.29
12	Alan		11/11/2004 9.48
23	Tim Murray		14/11/2004 4.56
1	**Req: Freestyle Script font**	Jerilyn	06/11/2004 1.27
9	**looking for a specific font .**	Jim	06/11/2004 3.28
16	Character		06/11/2004 7.00
27	Dick Margulis		06/11/2004 12.47
30	Jim		06/11/2004 19.41
20	..:: Pro ::..		06/11/2004 14.11
26	..:: Pro ::..		06/11/2004 14.19
14	Andreas Höfeld		06/11/2004 17.26
16	**I need the 'MT Extra' font.**	Chaos Master	06/11/2004 3.29
31	Thomas Ferguson		06/11/2004 7.04
18	Character		06/11/2004 7.05
15	Rez		06/11/2004 7.10
21	Chaos Master		06/11/2004 17.19
18	Rez		07/11/2004 21.12
7	**Font Party for Mac**	Anna	06/11/2004 9.47

Tav. 6. Una videata di Forté Agent (newsgroup: `comp.fonts`).

Per la comprensione del discorso si segue di solito una logica *bottom up,* inserendo il proprio intervento generalmente sotto il messaggio (o la parte di messaggio) a cui si desidera replicare[13]; è possibile accedere agli elementi che compongono la conversazione nell'ordine inverso a quello di inserimento. L'elemento inserito cronologicamente per ultimo, quindi, è il primo a cui si accede e man mano si risale, fino al primo elemento inserito, il capocatena che ha dato il via alla discussione. Questa struttura ben si presta a cogliere lo spostamento del focus attenzionale rispetto ad un cómpito strutturato gerarchicamente in "sottocómpiti" (terminologia di Clark - Schaefer 1989), collocati a livello di dettaglio via via crescente. Il primo elemento inserito nella struttura è costituito dallo spazio di conoscenza correlato al compito più generale; man mano

[13] Questo perlomeno è quando richiesto dalla Netiquette, nella convinzione che porre la replica in testa al nuovo messaggio, e dunque prima della parte quotata, porterebbe nel giro di alcuni messaggi alla non rintracciabilità dei singoli contributi ed a una maggiore difficoltà di comprensione. Di fatto, però, non è troppo raro trovare poster che hanno l'abitudine di riportare *tutto* il messaggio quotato sotto il proprio, uso probabilmente esportato dalle e-mail dove è quasi la norma (trattandosi di una comunicazione uno a uno, gli equivoci sono meno facili). Ma per una fenomenologia più dettagliata del quoting, cfr. § 1.3.1.

che si prendono in esame i suoi sottocómpiti, ed i loro componenti, si aumenta il livello di dettaglio inserendo nuovi elementi nella struttura, in modo tale che lo spazio di conoscenza relativo all'ultimo sottocómpito esaminato prima di quello corrente sia sempre in cima allo *stack*[14].

La schermata riportata in Tav. 6 chiarifica anche il processo di sviluppo del discorso, rendendo rintracciabile il suo "movimento referenziale" (Vater 2001), ovvero il modo in cui l'informazione si sviluppa da un enunciato all'altro[15], un movimento anche fisicamente visibile nella struttura di una gerarchia e nell'alternanza dei suoi *referentielle Domänen.*

Anche la tradizione tedesca, che pure ha visto il fiorire della maggior parte degli studi di matrice testuale relativamente sulla nozione di *referentielle Bewegung*, ha dovuto ad ogni modo scontrarsi con numerosi dubbi sul concetto di *testo* in ambiente di CMC. L'organizzazione non lineare degli ipertesti ha in particolare posto il problema della distinzione e definizione dell'ipertesto stesso, considerando un testo inteso come «Wortlaut, Folge untereinander in Zusammenhang stehender Sätze» (Naumann 2004) e l'ipertesto che invece prospetta enunciati presenti su diverse pagine e collegati tramite link – attivabili peraltro, non dimentichiamolo, solo a discrezione dell'utente, che quindi ha la facoltà di scegliere di volta in volta il percorso testuale che preferisce.

Le proposte presentate recentemente, soprattutto per l'analisi della comunicazione in chat, soccorrono il nostro lavoro a livello metodologico, in particolare nell'applicazione di modelli testuali a dati della CMC, ma non risolvono il nodo problematico che abbiamo già accennato: la specificità dei newsgroup, non solo nel loro collocarsi in un grado intermedio tra i poli scritto / orale, ma anche nelle dinamiche temporali e referenziali che li caratterizzano.

La natura composita dei newsgroup non pone solo problemi di adattamento di definizioni tradizionali ad una tipologia di testo particolare, ma dà origine ad alcuni interrogativi che coinvolgono i confini del testo stesso. La questione verte in particolare sull'opportunità di considerare "testo" tutto il thread o solo il singolo messaggio, od il singolo messaggio ma con tutti i quoting del thread. Ritorniamo in questo modo al meccanismo delle scatole cinesi: anche all'interno di un singolo thread, contenuto in un particolare genere di newsgroup, troviamo microtesti autonomi originati dal meccanismo del quoting e dal commento al testo ripreso. È d'altra parte vero che la coerenza rispetto al tema generale è data dal titolo del thread e dal fatto che tutti i post, seppur secondo uno svolgimento poco prevedibile dagli autori stessi, si attengono al tema centrale. I *topic shifts* sono in ogni caso dovuti a svolte tematiche che aprono sezioni di testo limitate solitamente a qualche scambio; inoltre, considerare i singoli post come testi completamente autonomi dal più vasto quadro del thread significherebbe ignorare la multiautorialità, la stretta connessione formale e contenutistica interna ai newsgroup, e la possibilità per i nuovi arrivati di riprendere parti di testo anche piuttosto in alto nella sequenza.

Potremmo allora considerare il singolo messaggio come il capitolo di un libro: un testo autonomo, ma inserito in un macrotesto che è il thread. I link ipertestuali costituirebbero poi rimandi "bibliografici" (apparati di solito non interni al libro). In realtà il riferimento abbastanza comune a fonti esterne, proprio tramite link – meno comunemente tramite allegati – modifica i termini del paragone. Nonostante quella che Gheno 2004, p.269, definisce «caparbia e spartana

---

[14] «In informatica, il termine *stack* o *pila* viene usato in diversi contesti per riferirsi a strutture dati le cui modalità d'accesso seguono una politica *LIFO* (Last In First Out), ovvero tale per cui i dati vengono estratti (letti) in ordine rigorosamente inverso rispetto a quello in cui sono stati inseriti (scritti). Il nome di questa struttura dati è infatti la stessa parola inglese usata, per esempio, per indicare una "pila di piatti" o una "pila di giornali", e sottende per l'appunto l'idea che quando si pone un piatto nella pila lo si metta in cima, e che quando si preleva un piatto si prelevi, analogamente, quello in cima (da cui la dinamica LIFO), anche se è possibile inserire o prelevare elementi anche dalla coda, infatti più in generale la pila è un particolare tipo di lista in cui le operazioni di inserimento ed estrazione si compiono dallo stesso estremo» (Wikipedia IT, *s.v.*).

[15] «Diese Entfaltung der Information von Äußerung zu Äußerung bezeichnen wir als referentielle Bewegung.» (Klein - Stutterheim 1987, p. 166).

testualità» (non vi sono disegni, immagini o musica), i newsgroup offrono la possibilità di uscire dal "gruppo/macrotesto" per visionare il file inserito in un altro scambio ed il documento può diventare il fulcro di scambi successivi – si vedano i passaggi, cit. come es. [10], in `it.comp.grafica.photoshop` sull'effetto cornice intorno ad una foto localizzabile in Internet:

[10]
```
In questa foto http://www3.photosig.com/viewphoto.php?id=442582
come è possibile ottenere l' effetto sfumato della cornice
bianca intorno al viso della modella ? Ovviamente con PS7 .
Ringrazio chiunque vorrà rispondermi .
```
it.comp.grafica.photoshop (NUNC-IT Photo).

Storrer nel trattare la coerenza negli ipertesti, riprende l'idea di Stutterheim 1997 per quanto riguarda la produzione di un testo: esiste una *Quaestio*, la domanda implicita a cui si deve dare una risposta in quel testo, «wird der Zusammenhang zwischen der thematischen Gesamtvorstellung, die der globalen Kohärenzbildung zugrunde liegt, und der Art des Textaufbaus mit Hilfe der Kategorie der Quaestio präzisiert» (Storrer 2000, p. 277).

```
From Tie.Fighter@libero.it Wed Jan 29 19:31:23 2003
Newsgroups: it.arti.scrivere
Subject: Re: Noce di burro
Date: Wed, 29 Jan 2003 19:31:23 +0100

Ignorando il Lato Oscuro della Forza Antonio Koch mise fine alla propria
esistenza con queste parole:

> > Fratello, a chi lo dici. Comunque stavo per cazziarti, per fortuna
> > che sono arrivato fino alla fi
> > ca, devi arrivare fino alla fi

Uaz, e pensavo foste persone serie... :]P

--

Lø¢k .·´¨¨))  -:¦:-
           ¸.·´  .·´¨¨))
        ((¸¸.·´   .·´  -:¦:-
         -:¦:- ((¸¸.·´  http://ow.too.it
                         ::Only Words::

    From grrrbau@hotmail.com Wed Jan 29 23:50:16 2003
    Newsgroups: it.arti.scrivere
    Subject: Re: Noce di burro
    Date: 29 Jan 2003 14:50:16 -0800
    Organization: http://groups.google.com/

    > Ignorando il Lato Oscuro della Forza

    [OT][OT][OT]
    Lo so che non ciazzeccaniente... ma levami una curiosità,
    in che rapporti stai con Star Wars?

    RP
```

Tav. 7. La ramificazione di un thread in `it.arti.scrivere` (da Forté Agent).

Data però la presenza di tale domanda nel *subject*, è pure davanti agli occhi di tutti i partecipanti ad un newsgroup un buon numero di casi in cui si esce fuori dal tema centrale del thread: sono gli stessi utenti a segnalarlo, attraverso l'indicazione [OT] (= *out of topic* od *off topic*). Per illustrare ciò abbiamo nella Tav. 7 integralmente riprodotto (direttamente dal newsreader, anziché mediato da NUNC) un segmento del thread da cui avevamo già estratto l'esempio [7]. Dall'intervento di "RP" nel campione di Tav. 7 nasce un nuovo dibattito, articolato in ben venti messaggi, completamente spostato sul tema Star Wars ed assolutamente dimentico della *Noce*

di burro che vediamo nel "Subject", ovvero il racconto proposto da uno dei partecipanti e successivamente commentato dagli altri iscritti di `it.arti.scrivere`. Se utilizziamo il concetto di *domini referenziali*, dobbiamo prendere in considerazione anche questi ulteriori scambi, che nella struttura del thread risultano una sorta di appendice. Possiamo quindi parlare di *coerenza locale*, quando prendiamo in esame la parti del testo (solo alcuni degli scambi) e di *coerenza globale* se consideriamo i costituenti inscritti in un quadro tematico più ampio (l'intero thread).

Ci sembra dunque che, nella delimitazione della dimensione testo-newsgroup, si possano tracciare dei confini abbastanza precisi ad inizio e fine thread[16], e non solo in relazione al singolo post od a un limitato gruppo di scambi, in modo da considerare un'unità di macroargomento dove sono possibili i *topic shifts* che descrivono parabole che si allontanano dal tema centrale per poi ritornarvi.[17]

Se poi vogliamo, sulla scorta di Maria-Elisabeth Conte, definire un testo in termini innanzitutto di coerenza (ciò che fa di un insieme di enunciati un testo, la *quidditas* del testo: Conte 1999/88), allora certamente dobbiamo esaminare il funzionamento dei newsgroup anche da questo punto di vista (cfr. oltre, § 1.3.1).

1.3.1 COERENZA E QUOTING. Coerenza e coesione nei messaggi di posta elettronica presentano, rispetto ad una lettera tradizionale, una caratteristica legata al medium: inglobano il messaggio a cui rispondono considerandolo a pieno titolo come co-testo (Fiorentino 2004; cfr. anche Garcea - Bazzanella 2002).

Il quoting, di cui abbiamo già delineato le caratteristiche fondamentali nel § 1.1, è un meccanismo attivato automaticamente dai newsreader, ma che necessita in fase di risposta ad un post anche della consapevole elaborazione dell'utente nell'attenta cernita di cosa "quotare" e cosa rimuovere dal messaggio, in modo da rendere il proprio post chiaro ed efficace, con le sole informazioni necessarie alla contestualizzazione della propria replica (visto inoltre che le risposte possono arrivare a distanza di giorni). La questione non è secondaria, se consideriamo il pullulare online di piccole guide sull'uso del quoting[18], così come il persistente richiamo all'interno dei thread a "quotare bene", rimproverando i newcomers che lasciano intatta la parte citata o che, viceversa, la eliminano del tutto.

Storrer 2000 parla a proposito del rapporto tra i testi che compongono un ipertesto di "sequenzializzazione del messaggio di risposta" (*Sequenzialisierung der Antwortnachricht*), di quel processo, cioè, che porta il fruitore a selezionare *un* percorso tra le possibilità offerte dal testo. Similmente la progressione tematica nei NG è fissata dalla successione dei messaggi che spesso si sovrappongono: passaggi di transizione vengono cancellati ed il testo si costruisce attraverso la selezione delle parti di testo salienti per gli interlocutori.

[16] Cfr. Marello 2007, p. 147: «Therefore it can be said that a NG thread is a text composed by many subtexts sharing the subject and having the same type of structure».

[17] La scelta di individuare l'identità del testo all'interno dei confini del thread è sottintesa anche dal trattamento informatico dei dati praticato nei NUNC: per evitare la ridondanza causata dal quoting (ossia, da diversa angolatura: per contenere il fenomeno del testo ripetuto, esiziale per indagini statistiche di interesse lessicografico) è stato usato un sistema di indicizzazione e script di filtraggio che ha eliminato i post più brevi all'interno di ogni thread, selezionando invece solo i messaggi più lunghi e più ricchi di citazioni, quelli cioè che più probabilmente contenevano quasi tutto il testo del thread (cfr. Barbera 2007 *i.s.*, e, nei dettagli, Casavecchia 2005, pp. 78-80).

[18] Un esempio tra tanti: «Un buon quoting è richiesto per due motivi: *a.* La maggior parte delle persone paga la connessione alla rete un tot al minuto, per cui pagare per dover scaricare un articolo che per la maggior parte è la copia integrale di qualcosa di già presente sul proprio PC non è gradevole. Se il lettore vuole avere un quadro più chiaro del contesto della discussione, si potrà sempre leggere il messaggio "padre". *b.* Un messaggio ben quotato è molto più comprensibile di un messaggio quotato male» (`http://digilander.libero.it/ifst/html/Quoting.html`).

Il quoting rappresenta un aspetto decisivo per lo studio della testualità in UseNet. Esso permette di cancellare alcuni mezzi connettivi fondamentali nei testi "tradizionali" come i meccanismi della ripetizione lessicale, dell'anafora e della ripresa del tema, nonostante nella scrittura in rete si registri uno sforzo (simile a quanto avviene nella conversazione ordinaria) di co-costruzione del senso e del testo con maggiore cooperazione.

Il riferimento alla costruzione della coerenza e della coesione in dipendenza della forte "contestualità" dà ragione di quelli che Andorno 2003, p. 158, definisce approcci "costruttivi", indicando che, attraverso il discorso e l'interazione, sono i parlanti stessi a costruire le categorie che ne regolano il funzionamento. Chi scrive nei newsgroup lascia le tracce che permettono agli altri partecipanti di proseguire il discorso, tracce di testualizzazione peculiari rispetto ad altri tipi di testo, ma che riflettono una forma particolare eppur conforme alle schematizzazioni dell'*Instruktionssemantik* (cfr. Conte 1999/88).

La teoria della sequenza di istruzioni accomuna testi scritti, parlati ed ibridi quali appunto i NG: tutte queste tipologie seguono comunque lo schema "apertura del tema (domanda/richiesta di informazione, ...), dibattito, eventuale spostamento del tema ecc.". Storrer 2002, pp. 9-11, valuta anche l'impatto della mancanza di sequenze testuali fisse («lack of a fixed text sequence») sulla costruzione e pianificazione della coerenza: il testo è sequenziale nel significato, non nella forma[19] in modo diverso per i NG, costituiti da nodi gerarchicamente ordinati, ma in cui il quoting spezza costantemente la linearità del discorso comunemente intesa. La nozione procedurale di coerenza ci pare nei due casi diversa: nelle pagine web è da vedersi nella sequela dei "movimenti" ipertestuali del lettore, coerenza costruita dal ricevente, *a parte subiecti*; in un NG, invece, la coerenza è costruita a partire da molteplici attanti: producente/i e ricevente/i contribuiscono in modo attivo all'organizzazione del testo ed alla creazione della coerenza testuale[20].

Una stessa informazione può inoltre essere ripresentata in vari modi, passando dalla ripetizione letterale, alla riformulazione parziale (implicante di solito l'impiego di proforme), alla riscrittura per mezzo di sinonimi lessicali, ed infine a forme intermedie fra la ripetizione e l'esplicitazione, quali la parafrasi ed in certi casi la riformulazione riassuntiva. Nei newsgroup la ripetizione sotto forma di quoting ad inizio messaggio ha soprattutto funzione di aggancio, ed implica la scomparsa dei tradizionali mezzi di ripresa, come congiunzioni e locuzioni congiuntive temporali, gerundi e participi. Compaiono poi forme intermedie tra il dispiegamento dell'informazione e la sua mera ripetizione: la parafrasi, ad esempio, e le riprese predicative, cioè quelle reiterazioni che, da un lato, fungono da incapsulatori e dall'altro qualificano ulteriormente questo evento.

Formalmente, il quoting può presentarsi in vari modi:

(1) Quoting del messaggio immediatamente precedente o di quello cui si risponde:
(2) Quoting di due o più messaggi per replicare a più di un intervento;
(3) Quoting di tutti i messaggi precedenti del thread;
(4) Quoting soltanto di una parte di un messaggio precedente, isolando una determinata frase per replicarvi in modo mirato;
(5) Quoting "spezzato": il messaggio si colloca in modo preciso in risposta a frasi estrapolate dai messaggi precedenti del thread.

Al di là di quanto già osservato (cfr. soprattutto note 8 e 13) le possibilità più sfruttate sono le cinque sopraelencate. La prima strategia è ovviamente la più semplice, e può essere rappresentata, nella sua forma più sintetica dall'esempio [11]:

---

[19] Questo vale per gli ipertesti intesi come da Todesco 2000 «ein Konglomerat von durch Hyperlinks verbundenen Textteilen auf einem Computer(verbund)».

[20] Non solo tramite la costante vigilanza sul quoting, ma anche adattando le normali strategie di distribuzione dell'informazione a un'alternanza *given-new* diversa da altre varietà testuali.

[11]
```
Subject : Re: Ottenere QUESTO effetto
" Valvola Digitale "  ha scritto :
> scommettiamo che è un drop shadow ? =)
Secondo me non lo saprai mai , visto che è quasi identico l'
effetto. Potresti scoprirlo solo dal valore cromatico dei pixel
interessati dall' " ombra " , che nel caso del drop shadow
standard sono affetti da un " multiply " , mentre nella mia
ipotesi si ha solo una sovrapposizione di livelli senza
interazione . Credo sia uno studio molto ardito ...
Mi è venuto istintivamente da optare per il glow , perché non è
molto intuitivo assegnare una distanza pari a 0 a un' ombra .
Sicuramente un principiante , applicando il drop shadow senza
troppe specifiche , si sarebbe trovato di fronte a un fenomeno
un po' diverso da quello cercato . Ho tentato solo di essere più
preciso , non prenderla come una provocazione .
Saluti ombreggiati ,
Alex
```
it.comp.grafica.photoshop (NUNC-IT Photo Uncut).

Le altre strategie sono più articolate; rinunciando per ragioni di spazio ad illustrare anche la (2) e la (3), esemplificheremo almeno la (4) e la (5) nella tavola 8[21] (cfr. *infra*), in cui sono illustrati tre post da un thread di `it.arti.musica.classica.mod`, dei quali il primo è il capocatena del thread, il secondo un esempio del quoting di tipo (4) ed il terzo uno del tipo (5).

L'esplicitazione, tra l'altro, di tutte le informazioni comporta anche una notevole riduzione delle operazioni di inferenza, dato che ogni scelta di strategia testuale è influenzata dalla citazione letterale dei frammenti di conversazione e poco è lasciato alla capacità dell'interlocutore di cogliere le informazioni implicite (ma la dimensione "sociale" fa sì che ci si possa riferire a thread precedenti e che solo gli appartenenti alla comunità virtuale possano cogliere i riferimenti impliciti).

2. I NUNC, PROBLEMI METODOLOGICI. Come già altrove accennato (cfr. Barbera ¶ 1, § 2.2.5) i vantaggi di corpora come i NUNC sono numerosi: a partire dalla rappresentatività in termini di lingua d'uso, fino alla grande abbondanza di varietà testuale e registro.

A fronte di indubbi vantaggi ed aspetti di interesse, il ricorso ad UseNet presenta anche alcuni svantaggi. Tra questi i più evidenti sono, in primo luogo, tutti quegli aspetti condivisi da gran parte della CMC che sono peculiarità del mezzo: emoticons, abbreviazioni, acronimi, ecc., già trattati da gran parte della letteratura (Storrer 2000, Schlobinski 2000, Fiorentino 2004, Gheno 2004, ecc.). Pur rappresentando un aspetto importante di un certo tipo di comunicazione, "sporcano" il testo (soprattutto dal punto di vista di un suo trattamento automatico). In secondo luogo, l'abbondanza di testo ripetuto, anche se a volte (quando effetto del quoting) testualmente rilevante e quindi "buono", è però dannoso per conteggi di frequenza e statistiche lessicali.

Altro problema lo pone la difficoltà di parametrizzazione del genere testuale: l'estrema varietà di tipologie (testi dialogici come quelli dei newsgroup di intrattenimento, testi argomentativi come quelli dei newsgroup di politica, testi narrativi e descrittivi, testi regolativi come possono essere le ricette nei newsgroup di cucina, ecc.), e soprattutto la non netta individuabilità della lingua dei newsgroup dal punto di vista dell'opposizione tra scritto e parlato, costituiscono infatti variabili difficili da tenere bene sotto controllo. Alcuni dei testi presentati precedentemente nelle tavole 7 e 8 ed negli esempi [1]-[11] dimostrano come siano presenti stili e

[21] Il carattere usato nella tavola è il Times anziché il Courier, quale dovrebbe invece essere, per ragioni di impaginazione.

registri estremamente idiosincratici anche all'interno di uno stesso newsgroup (o di uno stesso thread). Si tratta dunque di una questione che investe più livelli e che vede intersecarsi differenti piani di analisi.

From: "Erewhon" Newsgroups: it.arti.musica.classica.mod
Subject: Fusa Date: Tue, 25 Mar 2003 07:55:19 +0000 (UTC)

Nei libri di teoria non ho mai trovato menzione della nota da 1/128, che
però talvolta capita di incontrare (ad esempio nella penultima battuta
dell'Adagio della prima Sonata per violino solo di Bach). Tantissimi anni
fa, l'insegnante di solfeggio ci disse che questa nota si chiama "fusa",
mentre - secondo la garzantina - la fusa, nell'antica notazione mensurale,
rappresentava il valore immediatamente inferiore alla semiminima; nel '600
si trasformò nell'attuale croma. Probabilmente il valore della fusa è poi
ulteriormente "slittato" fino ad indicare, appunto, la nota da 1/128. Però
la garzantina di questo non dice niente. Qualcuno di voi ne sa qualcosa?
E, ancora più OT, è possibile inserire note da 1/128 con Finale?

Ciao,
Fabio

(4) From: "Valerio" Newsgroups: it.arti.musica.classica.mod
Subject: Re: Fusa Date: Tue, 25 Mar 2003 11:32:21 +0000 (UTC)

"Erewhon" ha scritto nel messaggio

> E, ancora più OT, è possibile inserire note da 1/128 con Finale?

Non ho sottomano Finale ma credo proprio di sì. Con Sibelius 2, che è
un programma di notazone straordinario, si possono addirittura
inserire note del valore di 1/128, 1/256, e 1/512!
Ciao
Valerio

(5) From: "Erewhon" Newsgroups: it.arti.musica.classica.mod
Subject: Re: Fusa Date: Tue, 25 Mar 2003 13:31:43 +0000 (UTC)

"Valerio" ha scritto nel messaggio

> > E, ancora più OT, è possibile inserire note da 1/128 con Finale?

> Non ho sottomano Finale ma credo proprio di sì.

Ale Redfiddler ha trovato la soluzione: Speedy entry: ctrl + 0 (© Ale
Redfiddler)

> Con Sibelius 2, che è
> un programma di notazione straordinario, si possono addirittura
> inserire note del valore di 1/128, 1/256, e 1/512!

Ehm... sullo "straordinario" stenderei un velo pietoso... :-)

(Mi spiego meglio: per una pubblicazione, mi è capitato nei giorni scorsi di
dover lavorare ad un file scritto originariamente con Sibelius. C'era una
valanga di "errori" sintattici che Finale non avrebbe mai permesso di fare,
se non con un durissimo lavoro di editing. Certo, colpa dell'autore del
file, ma, ripeto, con Finale il brano sarebbe stato molto più "corretto" da
un punto di vista sintattico).

Ciao,
Fabio

Tav. 8. Strategie diverse di quoting in alcuni post di un thread complesso in `it.arti.musica.classica.mod` (da Forté Agent).

Dal punto di vista del bilanciamento tematico interno la natura "democratica" dei NG assicura una distribuzione naturalmente omogenea degli argomenti, nel senso che è spontanea e "data": tanto da far pensare alla tassonomia dei newsgroup come ad una *folk taxonomy*[22], ed ai newsgroup stessi come ad una sorta di enciclopedia onnicomprensiva delle attività umane creata "dal basso", gerarchicamente ordinata, e che riveste a suo modo anche un'utilità sociale.

Tali aspetti sono di grande interesse per il linguista, sia egli un testualista, un lessicografo, un sociolinguista, un pragmalinguista, un esperto di linguistica antropologica, od altro ancora; ma si rivelano talvolta un'arma a doppio taglio, soprattutto qualora si voglia indagare un solo determinato aspetto del proteiforme e multifario mondo dei newsgroup. Alcune delle problematiche legate alla fisicità del testo, come le sporcature dovute ai set di caratteri od a frammenti vaganti di codice binario/html, la ridondanza data dal quoting, ecc.[23], sono state comunque risolte (o portate a livelli statisticamente accettabili) in fase di preparazione testi grazie a vari moduli di filtraggio, tokenizzazione e markuppatura (per i quali cfr. Casavecchia 2005, pp. 70-81).

		sequenza logica delle procedure di filtraggio
F1.		elaborando un NG alla volta, per ciascun posting si decide se è da considerare valido ed attendibile o di disturbo sottoponendolo a selezione tramite tre filtri anti-spam:
	F1a	filtro per la rimozione dei messaggi duplicati aventi lo stesso *subject*;
	F1b	filtro per la rimozione dei messaggi identici caratterizzati dal medesimo Message-ID;
	F1c	filtro per la limitazione del *cross-posting*: i posting spediti a troppi indirizzi di NG vengono eliminati;
F2.		all'interno dei posting rimasti si filtrano tutte le "impurità" residue (ad esempio, frammenti di materiale non testuale, stringhe di codice di programmazione o di formattazione);
F3.		per raccogliere solo messaggi aventi informazioni testuali "rilevanti"si selezionano solo i messaggi "pieni" e si scartano quelli troppo brevi;
F4.		rimangono ancora le ripetizioni del "testo quotato", che vengono eliminate in modo semplice ma efficace tramite la selezione del messaggio più lungo di ciascun thread.
		sequenza reale delle procedure di filtraggio
		F1c → F2 → F3 → F1a → F1b →F4

Tav. 9. Le procedure di pulizia nei NUNC (adattato da Casavecchia 2005, p. 81).

Obiezioni metodologiche potrebbero anche essere sollevate rispetto alla comparabilità interlinguistica dei sottocorpora che formano i NUNC da un punto di vista quantitativo e temporale: annate diverse e dimensioni diverse delle gerarchie scaricate possono essere infatti causa di ri-

[22] La *folk taxonomy*, tema inaugurato in antropologia da Durkheim 1912, e da tempo praticato dalla linguistica antropologica, è oggi ben presente soprattutto nella ricerca biologica (cfr. Berlin et alii 1973 e Healey 1993); ma non si vede perché dalle culture "primitive" in cui è più spesso studiato non possa essere riportato all'antropologia del vecchio Occidente ipercivilizzato: se ne sono infatti già avute interessanti e più generali applicazioni cognitive alla "antropologia della scienza" *tout court* (cfr. Atran 2001), e dai NUNC, crediamo, potrebbero venirne interessanti svolgimenti.

[23] Un ulteriore effetto di ridondanza è dato dalla ripetizione delle formule di saluto, che è un'altra delle peculiarità che distinguono i newsgroup dalle altre forme di comunicazione dialogiche in rete, poiché sono quasi sempre presenti in calce (*escatocollo*, termine di estrazione diplomatica con cui nelle *Guidelines* di NUNC abbiamo indicato le clausole conclusive dei post) ad ogni messaggio postato; Haase 1997, p. 78, stima che nel suo corpus di notizie la cifra delle formule di saluto sia solo nel 14% degli articoli analizzati.

sultati non direttamente confrontabili. In effetti per avere corpora comparabili interlinguisticamente, come coi NUNC si è cercato di fare, è spesso necessario parzialmente rinunciare ad uno dei due requisiti e, tra la coincidenza perfetta dei periodi di scarico e la comparabilità dimensionale, non v'è dubbio che è la seconda a dover essere privilegiata, contando inoltre sul fatto che alcuni mesi di differenza nella produzione ed acquisizione dei testi ben difficilmente possono comportare significative alterazioni della lingua sul piano sincronico[24].

Per questioni che riguardano più propriamente la metodologia di ricerca, invece, tre sono le strade più plausibili da seguire, a seconda di quali scopi il ricercatore si prefigge.

(1). Per studi di carattere generale sull'uso della lingua (cfr. ad es. Guil - Borreguero Zuloaga ¶ 18, Onesti - Squartini ¶ 15, ed Onesti ¶ 14) è utile interrogare i NUNC a tutto tondo nella versione standard, in cui sono state applicate tutte e quattro le procedure di filtraggio di Tav. 9; in tali versioni il contenimento del testo ripetuto è stato ottenuto semplicemente privilegiando in messaggio più lungo di ogni thread, e rinunciando all'integrità del thread medesimo.

(2). Il rumore di fondo, dato principalmente dal testo ripetuto, anche se fortemente abbattuto, resta però abbastanza rilevabile: se si intende condurre una ricerca puramente quantitativa, basata solo su dati e statistiche, come avviene ad esempio per le liste di frequenza nell'àmbito di ricerche terminologiche e lessicografiche, possono ancora emergere alcuni problemi residui (legati alla ridondanza e soprattutto alle ripetizioni) che potrebbero inficiare i risultati. È consigliabile allora creare delle stop lists di filtraggio in fase di eleborazione statistica, od usare in modo incrociato sottocorpora diversi (al modo illustrato poco oltre).

(3). Specularmente, laddove si vogliano invece osservare fenomeni testuali sulla scorta di quelli brevemente citati in questo articolo, la via da seguire sarà piuttosto quella di privilegiare l'integrità del thread a scapito della presenza di molto testo ripetuto: versioni apposite di alcuni sottocorpora sono pertanto state preparate, in cui agiscono solo i primi tre moduli di filtraggio menzionati nella Tav. 9, ma non il quarto (tra queste per ora il solo NUNC-IT Photo Uncut è anche disponibile online).

2.1 UN ESEMPIO: LE COLLOCAZIONI *ADJ-NOUN* NEI NUNC-UK. I NUNC, come si diceva, offrono la possibilità di lavorare su sottocorpora preordinati, già predisposti per l'interrogazione online: a seconda degli obiettivi della ricerca e del corrispondente interesse verso registri più o meno formali, o verso una terminologia specialistica anziché quotidiana, ci si potrà avvalere dei NUNC cucina, foto, motori, ecc., di dimensioni ridotte rispetto ai NUNC generici I e II, ma mirati ad aree di competenza linguistica differenziate; od addirittura, per interessi più testuali, alle versioni con thread non potati.

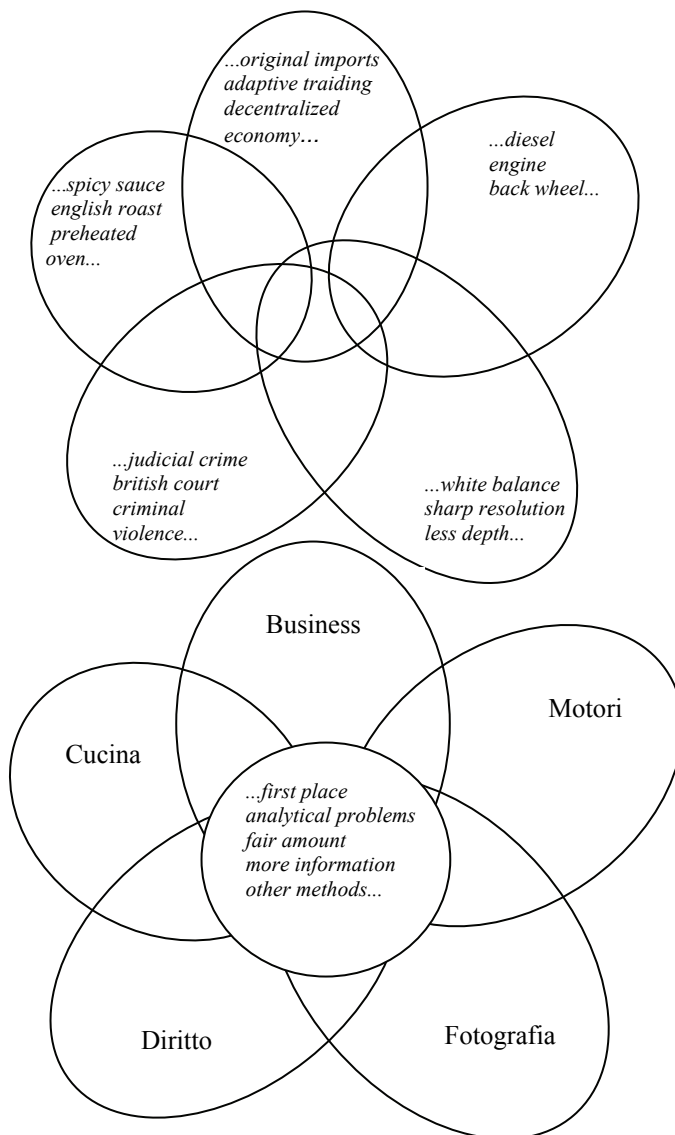
Al fine di meglio illustrare le potenzialità della combinazione di interrogazioni sui sottocorpora, esemplificandone una possibile strada metodologica per l'estrazione e l'analisi dei dati, riprendiamo ora parte del capitolo dedicato all'analisi delle collocazioni nei NUNC-UK dalla tesi di Sara Casavecchia, in questo senso esemplare. L'obiettivo che ci si proponeva è quello di verificare se il lessico, campionato nelle collocazioni aggettivo - nome (*adj-noun*), ed identificato come lessico specifico a partire dai corpora NUNC-UK specialistici (motori, cucina, photo, business e diritto), appartesse realmente al linguaggio specialistico di quel settore o meno.

A tal fine è stato anzitutto necessario estrarre[25], per ciascun corpus, la lista di coppie composte da aggettivo e nome, per poi procedere, tramite controlli incrociati, all'estrazione dalla

---

[24] Per quanto riguarda l'aspetto qualitativo, la correzione e la soluzione di eventuali errori quali ad esempio la presenza di "intrusi" di altre lingue all'interno di un corpus, rimandiamo a quanto accennato in Barbera 2007 *i.s.* sulla scorta di Grefenstette - Nioche 2000.

[25] La query di estrazione delle liste è definita come: `[pos = "JJ.*" & word="[a-z-]+" %cd] ([pos = ",|CC|JJ.*|RB"]*[pos = "JJ.*"])? [pos = "NN.*" & word="[a-z-]+" %cd];` in cui, secondo il sistema del Penn Treebank, il POS-tag "JJ" corrisponde all'aggettivo e "NN" al nome.

lista di ciascun corpus specialistico solo delle coppie che non comparissero nelle liste degli altri corpora specialistici. Quindi si sono messe a confronto le collocazioni specialistiche (*domain specific collocations*), isolate da ogni corpus settoriale, con la lista di collocazioni estratta dal corpus generale. Questo confronto è utile in primo luogo per determinare quali coppie di ogni lista sono realmente collocazioni specifiche; in secondo luogo per verificare quali coppie, tra quelle isolate come specifiche all'interno dei singoli corpora specialistici, vengono utilizzate nel corpus generale, e quindi nel linguaggio non specialistico, ed in che misura compaiono. In altre parole bisogna appurare che le liste estratte dagli specialistici come peculiarità siano davvero tali e non abbiano, invece, un numero comparabile di occorrenze anche nel corpus generale.

Tav. 10. Schema delle collocazioni specialistiche e generiche nei NUNC-UK (da Casavecchia 2005, pp. 111-112, Figg. 5.1 e 5.2).

«Per chiarire meglio – scriveva Casavecchia 2005, p. 110 – come è avvenuto il confronto tra le diverse liste, immaginiamo che ad ognuna di esse corrisponda un insieme. Ogni insieme contiene tutte le collocazioni *adj-noun* provenienti da un settore specifico (corpus specialistico), ma non tutte le collocazioni possono essere specialistiche, perché molte di esse vengono usate spesso in contesti generali (come *many thanks*) od in locuzioni avverbiali (come *same time*). [...] Il risultato del confronto degli insiemi è la scoperta di insiemi disgiunti, indipendenti, ovvero sottoinsiemi formati da coppie *adj-noun* che non compaiono negli altri insiemi e che rappresentano le liste di candidati ad essere termini "specialistici". L'estrazione di liste di coppie indipendenti ha l'obiettivo di restringere l'analisi a quelle che dovrebbero essere le *domain specific collocations*». La Tav. 10 rappresenta graficamente i cinque insiemi sottoposti a confronto incrociato: le sezioni più esterne rappresentano gli insiemi disgiunti, mentre quella interna, intersecante gli altri insiemi, è l'insieme congiunto, contenente collocazioni non specialistiche rintracciabili anche nel corpus generale (sono per lo più locuzioni avverbiali come *many times*, *only way*, ecc., o sintagmi nominali comuni quali *good idea*, *many thanks*, ecc.).

Come si può, inoltre, notare dalla tavola 11 (dove accanto ai valori assoluti sono dati anche quelli percentuali, per ovviare alla diversa dimensione dei vari corpora) vi è un elevato numero di coppie distinte *adj-noun* per ciascun corpus specialistico e vi è anche una spiccata prevalenza di collocazioni specialistiche (insiemi disgiunti) rispetto a quelle non-specialistiche (insieme congiunto), che conferma la grandissima varietà disponibile per indagini di carattere qualitativo.

settore	**coppie adj-noun specialistiche**		**coppie adj-noun non specialistiche**	**coppie totali**
Business	45.750	72,3%	17.509	63.259
Cucina	15.254	68,7%	6.935	22.189
Diritto	44.222	70,2%	18.808	63.030
Fotografia	6.780	60,7%	4.381	11.161
Motori	47.374	74,7%	16.025	63.399

Tav. 11. Valori assoluti e percentuali delle collocazioni specialistiche e generiche nei NUNC-UK (da Casavecchia 2005, p. 114, Tav. 5.3 e Fig. 5.3).

Estratte le liste di coppie *adj-noun* che non sono comuni a più corpora specialistici, e quindi circoscritti i termini che possono plausibilmente rientrare nelle *domain specific collocations*, l'obiettivo successivo era quello di confrontarle con la lista di collocazioni del corpus generale.

«Per permettere una simile comparazione bisognava ricercare le stesse collocazioni specialistiche», individuate con la procedura degli elementi disgiunti, «nella lista del corpus generale e confrontarne le occorrenze, per verificare che fossero davvero peculiarità del linguaggio specialistico e non avessero, invece, le stesse frequenze anche nel corpus generale. La frequenza, quindi, costituisce l'oggetto d'indagine centrale e lo strumento primario» (Casavecchia 2005, p. 115) della successiva fase di analisi: tramite il confronto tra i corpora specialistici e quello generale sono state selezionate le coppie ricorrenti in entrambe le liste e per ciascuna coppia sono state esaminate le frequenze, quella nel corpus generale e quella nel corpus specialistico (sottoinsieme del corpus generale). La differenza di questi valori indica la presenza di tale coppia in misura più o meno rilevante nel corpus specialistico. Ad esempio, se tale numero è pari a zero significa che la collocazione è peculiare del linguaggio specialistico e viene pertanto denominata "collocazione specialistica pura" poiché compare solo in esso. Se invece la differenza presenta un valore elevato, esprime il fatto che la collocazione non viene utilizzata solamente nel linguaggio specialistico ed è quindi una collocazione specialistica non esclusiva dei corpora specialistici. Al di là del calcolo della "percentuale di specificità" possibile per ogni coppia di

collocati, prendendo, ad esempio, in considerazione solo le "specialistiche pure", si ottengono i dati riportati nella Tav. 12. «In essa si può notare una caratteristica comune a tutti i settori: la quantità di collocazioni "pure" è tipicamente un ordine di grandezza inferiore alla quantità di collocazioni non esclusive del corpus specialistico, le quali a loro volta sono quantitativamente circa di un ordine di grandezza inferiore rispetto alle collocazioni specialistiche nel corpus generale» (Casavecchia 2005, p. 117).

settore	**collocazioni specialistiche**		
	"pure"	non esclusive del corpus spec.	nel corpus gen.
Business	10.368	76.077	437.954
Cucina	3.673	12.939	99.470
Diritto	17.797	142.149	529.969
Fotografia	2.098	5.974	29.895
Motori	23.939	68.962	221.548
tot.	**57.875**	**306.101**	**1.318.836**

Tav. 12. Cifre delle collocazioni specialistiche, da "pesare" e generiche nei NUNC-UK(da Casavecchia 2005, p. 117, Fig. 5.4).

Da questa prima analisi Motori e Diritto risultano i settori più tecnici: hanno più collocazioni "pure" degli altri (mediamente quasi 3 volte più degli altri, in percentuale il 6,5% contro 2,5%).

Partendo dai dati ottenuti è possibile ricavare la percentuale delle diverse tipologie di collocazione sul totale delle coppie *adj-noun* nel corpus generale, ossia "pesare" la terminologia specialistica (la pura e la non esclusiva) in rapporto alle frequenze delle coppie estratte dal corpus generale. Ed emerge, tra l'altro, che «la percentuale relativa alle collocazioni "pure", o *domain specific collocations*, è minima (inferiore all'1%); quella delle collocazioni specialistiche non esclusive dei corpora specialistici è pari al 4,2%, ma la somma delle collocazioni specialistiche all'interno dei settori specialistici è di circa 4 volte inferiore al numero delle collocazioni specialistiche nel corpus generale» (Casavecchia 2005, p. 118), citcostanza che potrebbe essere da imputare al fatto che nei generici il ventaglio di soggetti discussi è molto più vasto che negli specifici.

E l'ottima ricerca della Casavecchia non si ferma qui; ma ai nostri scopi, che sono poi solo quelli di illustrare un punto metodologico, può in questa sede bastare.

3. CONCLUSIONI. Il *case study* parzialmente riportato, rivolto ad una terminologia settoriale, è eloquente rispetto alle capacità potenziali offerte dai NUNC a partire dall'elevato tasso di gerarchie specialistiche, che convogliano esperti della materia in una comunità virtuale in cui la comunicazione tra gli stessi membri, pur tecnica, è diversa da qualsiasi altra *Textsorte* per immediatezza comunicativa, costruzione comune della coerenza, ripresa testuale chiaramente ancorata alla *quaestio* del thread.

La coesistenza di tali fattori, la loro pregnanza testuale, e soprattutto il peculiare svolgimento del discorso quale *continuum* nella discontinuità del quoting, rendono i newsgroup un'area di ricerca linguistica stimolante, in cui dati autentici ed adeguati strumenti di interrogazione, come quelli forniti dai NUNC, consentono di indagare complesse varianti della lingua moderna.

BIBLIOGRAFIA.

ALLORA
2003 Adriano Allora, *È scritto o parlato?,* in "Italiano & Oltre" I (2003) 14-18.
2005 Adriano Allora, *A Tentative Typology of Net Mediated Communication*, comunicazione presentata alla *Corpus Linguistics 2005 Conference, Birmingham July 14-17 2005*, disponibile online alla pagina `http://www.corpus.bham .ac.uk/PCLC/`
i.p. Adriano Allora, *Per una tipologia della comunicazione mediata dalla rete. Variazione diamesica generale*, in corso di stampa in "Bollettino dell'Atlante Linguistico Italiano".

ANDORNO
2003 Cecilia Andorno, *Linguistica testuale. Un'introduzione,* Carocci, Roma, 2003 "Università" 519.

ANDROUTSOPOULOS
2003 Jannis K. Androutsopoulos, *Online Gemeinschaften und Sprachvariation*, in "Zeitschrift für germanistische Linguistik" XXXI (2003) 173-197.

ATRAN
2001 Scott Atran, *Folk Biology and the Anthropology of Science. Cognitive Universals and Cultural Particulars*, preprint to "Behavioral and Brain Sciences", deposited on 30th April 2001, online at `http://www.bbsonline.org/Preprints/OldArchive/bbs.atran.html`.

BARACCO
2002 Alberto Baracco, *La comunicazione mediata dal computer*, in BAZZANELLA 2002, pp. 253-267.

BARBERA
2007 *i.s.* Manuel Barbera, *I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo*, in "Cuadernos de filología italiana" XIV (2007) 11-32, in corso di stampa.

BAZZANELLA
1994 Carla Bazzanella, *Le facce del parlare: un approccio pragmatico all'italiano parlato*, La nuova Italia, Scandicci, 1994 "Biblioteca di Italiano e oltre".
2002 *Sul dialogo. Contesti e forma di interazione verbale*, a cura di Carla Bazzanella, Milano, Guerini, 2002.

BERLIN et alii
1973 Brent Berlin - Dennis E. Breedlove - Peter H. Raven, *General Principles of Classification and Nomenclature in Folk Biology*, in "American Anthropologist" VII (1973) 214-242.

BERRUTO
1985 Gaetano Berruto, *Per una caratterizzazione del parlato: l'italiano parlato ha un'altra grammatica?*, in HOLTUS - RADTKE 1985, pp. 120-153.

BdT → PILLET - CARSTENS 1933

BRINKER
2001 Klaus Brinker, *Linguistische Textanalyse - Eine Einführung in Grundbegriffe und Methoden*, Erich Schmidt Verlag, Berlin, 2001.

CASAVECCHIA
2005 Sara Casavecchia, *Progettazione ed implementazione di corpora di lingua inglese basati sui newsgroups*, Tesi di laurea, Facoltà di lingue e letterature straniere, Università di Torino, 2004-2005.

CLARK - SCHAEFER
1989 Herbert H. Clark - Edward F. Schaefer, *Contributing to discourse*, in "Cognitive Science" XIII (1989) 259-294.

CONTE
1999/88 Maria-Elisabeth Conte, *Condizioni di coerenza*, Alessandria, Edizioni dell'Orso, 1999. Nuova edizione, con l'aggiunta di due saggi a cura di Bice Mortara Garavelli, di Maria-Elisabeth Conte, *Condizioni di coerenza. Ricerche di linguistica testuale*, Firenze, La Nuova Italia Editrice, 1988 "Pubblicazioni della Facoltà di Lettere e filosofia dell'Università di Pavia" 46.

CONTINI
1936 Gianfranco Contini, *Per la conoscenza di un sirventese di Arnaut Daniel*, in "Studi Medievali" ns. IX (1936), pp. 223-231.

CRYSTAL
2001 David Crystal, *Language and the Internet*, Cambridge University Press, 2001 [2006_2].

DURKHEIM
1912/2003 Émile Durkheim, *Les formes élémentaires de la vie religieuse: le système totémique en Australie*, Paris, F. Alkan, 1912. [riedizione moderna: Paris, PUF, 2003 "Quadrige"]

EUSEBI
1984 Arnaut Daniel, *Il sirventese e le canzoni*, a cura di Mario Eusebi, Milano, All'insegna del pesce d'oro, 1984.

FALKENHAGEN - LANDE
1997 Lena Falkenhagen - Svenja Landje, *Newsgroups im Internet*, [Hannover], Universität Hannover - Seminar für deutsche Literatur und Sprache, Wintersemester 1997-98 "HS: Sprache und Kommunikation @ Internet", online alla pagina http://www.mediensprache.net/networx/networx-1/inhalt.htm.

FEENBERG
1989 Andrew Feenberg, *The written world: On the theory and practice of computer*, in MASON - KAYE 1989, pp. 22-39.

FELDWEG - KIBIGER - THIELEN
1995 Helmut Feldweg - Ralf Kibiger - Christine Thielen, *Zum Sprachgebrauch in deutschen Newsgruppen*, in "Osnabrücker Beiträge zur Sprachtheorie" L (1995) 143-154, disponibile anche online http://www.sfs.uni-tuebingen.de/Elwis/news.ps.

FIORENTINO
2005 Giuliana Fiorentino, *Così lontano, così vicino: coerenza e coesione testuale nella scrittura in rete*, in KORZEN 2005, solo nel Cd-rom allegato.

FIORI
2004 Silvia Fiori, *An analysis of linguistics newsgroups through their paratexts*, in "Rassegna italiana di linguistica applicata" XXXVI (2004)$^{2-3}$ 67-81.

FIX - ADAMZIK - ANTOS - KLEMM

2002 *Brauchen wir einen neuen Textbegriff?,* Ulla Fix - Kirsten Adamzik - Gerd Antos - Michael Klemm (Hrsg.), Peter Lang, Frankfurt am Main, 2002.

GARCEA - BAZZANELLA

2002 Alessandro Garcea - Carla Bazzanella, *Discours rapporté et Courrier Electronique*, in "Faits de Langues" XIX (2002) 233-246.

GHENO

2004 Vera Gheno, *Prime osservazioni sulla grammatica dei gruppi di discussione telematici di lingua italiana*, in "Studi di Grammatica Italiana" XXII (2004) 267-308.

2005 Vera Gheno, *Mini-compendio sulla lingua dei newsgroup*, in "Italiano Accessibile", online alla pagina `http://www.italianoaccessibile.it/detail.asp?idn=2871`.

GLIOZZO - STRAPPARAVA

2004 Alfio Gliozzo - Carlo Strapparava, *Domain models for lexical semantics*, presentato durante l'*International Colloquium "Word Structure and Lexical Systems: models and applications", 16-17 dicembre 2004, Pavia*, disponibile online alla pagina `http://tcc.itc.it/people/gliozzo/talks/Domain_Models_Pavia.pdf`.

GREFENSTETTE - NIOCHE

2000 Gregory Grefenstette - Julien Nioche, *Estimation of English and non-English Language Use on the WWW*, in *Proceedings of RIAO 2000, 6th Conference: Content-Based Multimedia Information Access, Paris, April 12-14, 2000*, Paris, Collège de France, pp. 237-246, disponibile online come Arxiv preprint cs.CL/0006032 all'URL `http://arxiv.org/ftp/cs/papers/0006/ 0006032.pdf`.

GRICE

1989/93 Paul Grice, *Logica e conversazione. Saggi su intenzione, significato e comunicazione*, traduzione italiana di Giorgio Moro, Bologna, il Mulino, 1993 [edizione originale: Paul Grice, *Studies in the way of words*, Cambridge (Mass.) - London, Harvard University Press, 1989].

GÜNTHER - LUDWIG

1994 *Schrift und Schriftlichkeit | Writing and Its Use*, herausgegeben von | edited by Hartmut Günther und | and Otto Ludwig, Berlin, de Gruyter, 1994.

HAASE et alii

1997 Martin Haase - Michael Huber - Alexander Krumeich - Georg Rehm, *Internetkommunikation und Sprachwandel*, in WEINGARTEN 1997, pp. 51-85.

HAUBEN - HAUBEN

1997 Michael Hauben - Ronda Hauben, *Netizens. On the History and Impact of Usenet and the Internet*, Foreword by Thomas Truscott, Los Alamitos (CA), Wiley-IEEE Computer Society Press, 1997, disponibile anche online alla pagina `http://www.columbia.edu/~rh120/`.

HEALEY

1993 Christopher Healey, *Folk Taxonomy and Mythology of Birds of Paradise in the New Guinea Highlands*, in "Ethnology" XXXII (1993) 19-35.

HINRICHS et alii

1995 *Abschlußbericht [zu ELWIS Projekte]*, Projektleiter Prof Dr Erhard W. Hinrichs, Mitarbeiter Helmut Feldweg, Marie Boyle-Hinrichs und Ralf Hauser, PS file online `http://www.sfs.uni-tuebingen.de/Elwis/abschlussbericht.ps`.

HOLTUS - RADTKE
1985 *Gesprochenes Italienisch in Geschichte und Gegenwart*, herausgegeben von Günther Holtus und Edgar Radtke, Tübingen, Narr, 1985 “Tübinger Beiträge zur Linguistik” 252.

KALLMEYER
2000 *Sprache und neue Medien*, herausgegeben von Werner Kallmeyer, Berlin u.a, de Gruyter, 2000 “Institut für deutsche Sprache” Jahrbuch 1999.

KLEIN - STUTTERHEIM
1987 Wolfgang Klein - Christiane von Stutterheim, *Quaestio und referentielle Bewegung in Erzählungen*, in “Linguistische Berichte” CIX (1987) 163-183.
1992 Wolfgang Klein - Christiane von Stutterheim, *Textstruktur und referentielle Bewegung*, in “Zeitschrift für Literaturwissenschaft und Linguistik ”LXXXVI (1992) 67-92.

KOCH - ÖSTERREICHER
1994 Peter Koch - Wulf Österreicher, *Funktionale Aspekte der Schriftkultur*, in GÜNTHER - LUDWIG 1994, pp. 587-604.

KORZEN
2005 *Lingua, cultura e intercultura: l’italiano e le altre lingue. Atti del VIII Silfi, Societa Internazionale di Linguistica e Filologia Italiana (Copenaghen, 22-26 giugno 2004)*, a cura di Iørn Korzen, Copenaghen, Samfundslitteratur Press, 2005 “Copenhagen Studies in Language” 31.

KORZEN - LUNDQUIST
2007 *Comparing Anaphors between Sentences, Texts and Languages. Proceedings of the international symposium held at the Copenhagen Business School, September 1st-3rd 2005*, edited by Iørn Korzen and Lita Lundquist, Frederiksberg, Samfundslitteratur Press, 2007 “Copenhagen Studies in Language” 34.

LENKE - SCHMITZ
1995 Nils Lenke - Peter Schmitz, *Geschwätz im ‘Globalen Dorf’ - Kommunikation im Internet*, in “Osnabrücker Beiträge zur Sprachtheorie” L (1995) 117-141.

MARELLO
2007 Carla Marello, *Does Newsgroups “Quoting” Kill or Enhance Other Types of Anaphors?*, in KORZEN - LUNDQUIST 2007, pp. 145-157.

MASON - KAYE
1989 *Mindweave: Communication,computers and distance education*, edited by Robin Mason and Anthony Kaye, Oxford, Pergamon Press, 1989.

MITCHELL
1997 Tom M[ichael] Mitchell, *Machine Learning*, New York, McGraw-Hill, 1997 “McGraw-Hill Series in Computer Science”.

NAUMANN
2004 Anja Naumann, *Wissenserwerb und Informationssuche mit Hypertexten: Die Bedeutung von Strukturierung, Navigationshilfen und Arbeitsgedächtnisbelastung*, PhD Dissertation, disponibile online alla pagina `http://archiv.tu-chemnitz.de/pub/2004/0117/data/Naumann.pdf`

PETŐFI
2004 Petőfi [*sic*] János S[ándor], *Scrittura e interpretazione. Introduzione alla testologia semiotica dei testi verbali*, Roma, Carocci Editore, 2004 “Università” 613.

PILLET - CARSTENS
1933 Alfred Pillet - Henry Carstens, *Bibliographie der Troubadours*, Halle a. S., Max Niemeyer, 1933 "Schriften der Königsberger Gelehrten Gesellschaft. Sonderreihe" 3.

RAINER - STEIN
2003 *I nuovi media come strumenti per la ricerca linguistica*, a cura di Franz Rainer ed Achim Stein, Peter Lang, Frankfurt a.M., 2003.

RUSCH - SCHMIDT
2000 *Konstruktivismus in Psychiatrie und Psychologie*, herausgegeben von Gebhard Rusch und Siegfried J. Schmidt, Frankfurt, Suhrkamp, 2000 "Delfin" 1998/99.

SCHLOBINSKI
2000 Peter Schlobinski, *Anglizismen im Internet*, Hannover, Mediensprache.net, 2000 "Network" 14, online alla pagina `http://www.mediensprache.net/networx/networx-14.pdf`

SCHOLZ
2003 Arno Scholz, *Comunicazione giovanile in rete. Una mailing list italiana dedicata alla cultura hip-hop*, in RAINER - STEIN 2003, pp. 117-139.

SCHÖNEFELD
2001 Tim Schönefeld, *Bedeutungskonstitution im Hypertext*, Hamburg, Mediensprache.net, 2001 "Network" 19, online alla pagina `http://www.websprache.net/networx/docs/networx-19.pdf`

SPITZER
1922/2007 Leo Spitzer, *Italienische Umgangssprache*, Bonn, Kurt Scroeder, 1922. Versione italiana: *Lingua italiana del dialogo*, a cura di Claudia Caffi e Cesare Segre, traduzione di Livia Tonelli, Milano, il Saggiatore, 2007.

STORRER
2000 Angelika Storrer, *Was ist "hyper" am Hypertext?*, in KALLMEYER 2000, pp. 222-252, disponibile anche online alla pagina `http://www.deutsch.fb15.uni-dortmund.de/Members/040_forschung/Publikationen/pub-as_html`.
2002 Angelika Storrer, *Coherence in Text and Hypertext*, in "Document Design" III (2002)[2] 156-168, disponibile anche online alla pagina `http://www.deutsch.fb15.uni-dortmund.de/Members/040_forschung/Publikationen/pub-as_html`

THOME
2001 Matthias Thome, *Semiotische Aspekte computergebundener Kommunikation*, Saarbrücken, Mediensprache.net, 2001 "Network" 20, online alla pagina `http://www.mediensprache.net/networx/networx-20.pdf`

TODESCO
2000 Rolf Todesco, *Hypertext oder Was heisst Konstruktion im konstruktivistischen Diskurs?*, in RUSCH - SCHMIDT 2000, stw 1503, disponibile online alla pagina `http://www.hyperkommunikation.ch/todesco/publikationen/T_delf2.htm`.

VATER
2001 Heinz Vater, *Einführung in die Textlinguistik. Struktur und Verstehen von Texten*, 3. Auflage, Wilhelm Fink Verlag, München, 2001 "UTB für Wissenschaft".

WEINGARTEN
1997 *Sprachwandel durch Computer*, herausgegeben von | edited by Rüdiger Weingarten, Westdeutscher Verlag, Opladen, 1997.

CORPORA E SITI DI RIFERIMENTO.

20 Newsgroups	http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.data.html.
ELWIS	http://www.sfs.uni-tuebingen.de/Elwis/
Google Groups	http://groups.google.it/.
Arianna Usenet	http://arianna.libero.it/usenet/.
Big8	http://www.big-8.org/.
Forté Agent	http://www.forteinc.com/main/homepage.php.
NUNC	http://www.bmanuel.org/projects/ng-HOME.html.
Usenetportal	http://www.usenetportal.com.
Wikipedia	http://en.wikipedia.org/wiki/.
Wikipedia IT	http://it.wikipedia.org/wiki/.

CRISTINA ONESTI
(Università di Torino)

14. "Niusgrup"… si scrive così?[1]
Grafie in rete.

ADR. *Voce dal sen fuggita*
Poi richiamar non vale:
Non si trattien lo strale,
Quando dall'arco uscì.
Pietro Metastasio, *Ipermestra*, II.1.

0. INTRODUZIONE. L'accesso a dati di lingua scritta tratti da gruppi di discussione a libero accesso ("newsgroup") ha consentito la presente analisi *corpus-assisted* che si propone di indagare fenomeni di riflessione metalinguistica nella scrittura digitale in rete.

Il corpus di base adottato è l'insieme dei NUNC (*Newsgroups UseNet Corpora*) di lingua italiana (c. 237 milioni e mezzo di parole i soli "generici": cfr. in questo volume Barbera ¶ 1, Tav. 2), interessanti in quanto rappresentativi di un italiano mediato dal web: si tratta di un tipo di comunicazione scritta ed offline (cfr. Barbera 2007 *i.s.* e qui Corino ¶ 13 e Barbera ¶ 1, § 2.2.5), ma con un grado di interattività simile a quello della comunicazione faccia a faccia.

Ripetutamente osservato nella letteratura sulla lingua in rete è il richiamo alla lingua orale, che però ha finora subito da parte dei linguisti un'attenzione in parte distorta, spesso legata soltanto «alla presenza di interiezioni, ideofoni, emoticons, espressioni gergali o volgari (certamente tentativi di rendere alcuni tratti del discorso orale, cosa che ha per l'italiano valenza particolare, visto lo sviluppo diacronico diversificato che lingua orale e scritta hanno seguito), ma che colgono solo una dimensione stilistico-espressiva superficiale, forse importante ma non esaustiva» (Corino ¶ 13, § 1.1). Tralasceremo in questa sede la discussione sulla natura eterogenea dei testi oggetto della CMC, senza entrare nella *querelle* sulla maggiore propensione di tali produzioni verso il polo della lingua scritta o quello della lingua parlata (cfr. Allora 2003 e Corino cit.), notando però che, per quanto la lingua di chat, forum, newsgroup, ecc. possa tendere alla riproduzione di caratteristiche dell'oralità, rimangono pur sempre alcuni punti fissi che la àncorano alla scrittura.

Perché, di fatto, gli utenti *scrivono*. E questo implica che essi debbano riprodurre graficamente e non acusticamente un significato da veicolare ed addentrarsi in riflessioni metalinguistiche che investono il piano della grammaticalità della lingua e, soprattutto, il piano del significante delle parole. Non sempre tale riproduzione, anche nei madrelingua, avviene senza difficoltà: troviamo nei testi di forum online e newsgroup errori di vario tipo, spesso di battitura, legati alla velocità di scrittura nel web, ma anche riflessioni su termini a frequenza medio-alta e comuni dubbi ortografici, resi pubblici con la formulazione di un'interrogativa diretta del tipo "si scrive così?" / "come si scrive?", che ritroviamo sùbito dopo il termine dubbio e spesso tra parentesi, a mo' di inciso. Si è voluta verificare la natura dei contesti in cui lo scrivente, inserendo questo tipo di parentetiche, riflette esplicitamente sulla forma grafica delle parole.

[1] Una bozza preliminare del presente contributo è stata presentata insieme ad Elisa Corino, cui va il mio ringraziamento, in forma di poster dal titolo *"Si scrive così?" Difficoltà di ricezione/produzione del lessico tecnico-scientifico in un corpus di Newsgroups*, al Convegno Internazionale *L'Università: ponte tra Scienza e Società,* 15 e 16 settembre 2006, organizzato da Agorà Scienza, Università degli Studi di Torino, nella sessione "Attività di diffusione della cultura scientifica rivolte ad un pubblico non specialistico".

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 253-270.

1. LA COMUNITÀ DEI NEWSGROUP. Alcune brevi linee descrittive delle interazioni linguistiche nei newsgroup rendono ragione della peculiare realizzazione linguistica analizzata e legittimano uno studio a partire da corpora come i NUNC.

Diversamente dal contesto comunicativo creato, per esempio, dalla corrispondenza via e-mail, gli utenti di newsgroup partecipano ad una vera e propria comunità, con regole e tradizioni interne; aspetto da non sottovalutare in termini di ricaduta sul linguaggio per il concetto di *vicinanza* – pur virtuale in questo caso – tanto fondante nella riflessione sull'oralità secondaria (cfr. Ong 1982/86). Quella «mistica partecipatoria, il senso della comunità, la concentrazione sul momento presente e l'utilizzazione di formule» stimate da Ong cit. (p. 191 trad. it.), si riverberano nella percezione di spazio comune condivisa dai partecipanti ad un newsgroup, con interessi che li avvicinano e forti relazioni che segnano la cifra dell'appartenenza ad una comunità in cui identificarsi, in quanto rete di comunicazione autodefinita da scopi comuni (come il "villaggio globale" di McLuhan - Fiore 1968). Non per nulla vige all'interno della comunità una serie di regole da rispettare (la *netiquette*); i rapporti tra i veterani vengono assimilati in alcuni appellativi a quelli fraterni; ritroviamo persino una dimensione storica nel riferirsi a messaggi o scambi vecchi di mesi ma ben noti alla maggior parte degli utenti. Tale senso di appartenenza ad un gruppo circoscritto induce una certa libertà di espressione, che avvicina i testi dei NUNC ad alcuni aspetti della lingua orale.

In questa direzione si rivelano interessanti le osservazioni che Berruto faceva già nel 1985 sull'italiano parlato: il parlato avrebbe la stessa grammatica dello scritto, ma più liberalizzata e più focalizzata sull'emittente, il quale in base al contesto reinterpreta le regole del sistema linguistico, valendosi di una vicinanza spaziale assente nella lingua scritta (cfr. anche Koch - Österreicher 1985 e 1994). La particolarità dei newsgroup in tale prospettiva consiste proprio nel presentare una vicinanza analoga: non una reale prossimità spaziale, bensì quella virtuale di cui sopra, una contestualità che manca delle circostanze extralinguistiche in cui i parlanti sono abitualmente immersi nella "normale" comunicazione orale, ma accorcia comunque la distanza ontologica fra i soggetti coinvolti. Ritroviamo infatti nei newsgroup una liberalizzazione della lingua e focalizzazione sul parlante (in questo caso *scrivente*) tipici del parlato, con forme di quell'egocentrismo postulato da Berruto in termini di salienza emotiva e di discorso centrato sugli attanti (cfr. Berruto 1985, p. 143).

Si noti inoltre come alcuni elementi siano da considerarsi legati al modo, concretamente diverso, in cui si svolge la sequenza dialogica all'interno dei newsgroup ed al conseguente impatto che il mezzo ha sull'utente (quasi un'estensione degli organi di senso, *à la* McLuhan, in termini di influenza sulla comunicazione). Come osserva Scholz 2003, p. 127, «i generi testuali elettronici costituiscono uno spazio in cui l'azione di una norma standard viene meno. Naturalmente il rispetto della norma scritta dipende di gran lunga dallo scrivente e molto meno dal mezzo. Comunque il carattere di semipermanenza dei generi digitali sembra che favorisca […] una produzione scritta poco incline a rispettare le norme vigenti per testi scritti».

L'osservazione non rende però del tutto giustizia ai numerosi utenti che dimostrano non solo un'elevata competenza della terminologia specialistica a livello lessicale, ma anche una notevole cura testuale nello stilare post a scopi argomentativi od informativi. Senza tentare una classificazione esauriente del mondo dei newsgroup, troppo variegato nelle sua manifestazioni sia stilistiche sia tematiche (cfr. ancora qui Corino ¶ 13), possiamo però prendere atto di una condivisa immediatezza degli interlocutori: nella creatività contenutistica, ma ancor più in quella espressiva e stilistica, forgiata altresì da giochi linguistici che si nutrono frequentemente di acuta coscienza metalinguistica. Le stesse parentetiche analizzate in questo contributo occorrono talvolta a mo' di burla, di vezzo linguistico: pur senza negare un'effettiva ignoranza sulla grafia da cui sono generate, l'esigenza personale di dimostrare la consapevolezza dell'errore convive con lo scherzo, che lascia in ultima analisi inalterata e senza verifica la grafia in questione.

La percezione di appartenenza ad una comunità dà inoltre adito a mosse volte alla gestione della "faccia" (cfr. Andorno 2003, pp. 170-176): la volontà di mantenere un ruolo o semplicemente di non fare brutta figura all'interno del gruppo porta lo scrivente ad ammettere quantomeno la coscienza del dubbio ortografico, utilizzando strategie di mitigazione tese a costruire la propria identità sociale all'interno del gruppo di discussione (cfr. Goffmann 1964, Caffi 2001), ribadendola costantemente nel corso dello scambio comunicativo.

2. RIFLETTERE SULLA GRAFIA. Durante il processo di scrittura lo scrivente è sempre, più o meno inconsciamente, coinvolto nella riflessione ortografica, sia che si tratti di scrittura tradizionale su supporto cartaceo, sia di comunicazione online. L'interesse qui privilegiato alle forme di trasmissione in rete ed ai NUNC è legato alle caratteristiche analizzate nel § 1: il controllo più lasso rispetto a situazioni di corrispondenza cartacea e la presenza di formulazioni "rilassate", non necessariamente dipendenti da una rigida correttezza grammaticale, derivano proprio dalla percezione di una comunità, evidente nella conclusione del seguente esempio:

[1] Quello dell' immigrazione è un falso problema , cosa ? sono quasi tutti clandestini senza nessun controllo !! molti sono avanzi di galera ! Questa è la versione attuale dell' apartaid (non **si scrive così** ma capiamoci) . NUNC-IT Generic I.

Nei casi di dubbio ortografico si assiste pertanto nei NUNC all'utilizzo di formule come "si scrive così" / "come si scrive" che lasciano inalterati e non verificati a livello ortografico i termini su cui si manifesta il dubbio. L'adozione di strategie di evitamento non pare contemplata, se non in casi poco frequenti:

[2] Ciao a tutti , sapete dirmi in quali numeri di Urania Classici sono comparsi i romanzi del cislo dell' inquisitore di Evangelisti (e visto non ricordo **come si scrive** , ometterò di farlo) ? NUNC-IT Generic I.

Nella maggior parte delle occorrenze di "si scrive così" / "come si scrive" la pianificazione della frase non viene scompaginata, mantenendo il termine pur nell'incertezza ortografica, ma manifestando apertamente tale incertezza agli altri partecipanti.

Anche quando il soggetto trattato è marcatamente specialistico e subentrano pareri di esperti od appassionati con contributi linguisticamente controllati, accade che anch'essi abdichino senza riserve ad alcune norme di scrittura o di formalità del messaggio. Se la sorveglianza dello scritto è forse un metodo per veicolare nel newsgroup maggiore ufficialità e serietà, che valorizzi esteriormente l'autorevolezza dell'informazione veicolata, ciò non esclude una frequente commistione di considerazioni tecniche con linguaggio colloquiale, scomposto, permeato di interiezioni e termini volgari, nonché talora di dubbi ortografici.

Nell'estrazione delle interrogative "si scrive così?" / "come si scrive?", si sono voluti mettere a fuoco i termini sulla cui forma scritta gli scriventi indugiano, rallentando il processo di scrittura e chiedendosi (a se stessi prima ancora che, in modo pubblico, agli altri interlocutori) quale sia l'esatta grafia di una parola.

La «voce dal sen fuggita» evocata in epigrafe non si può certo richiamare nel parlato, ma il dato interessante è che, là dove invece si potrebbe, ovvero nella lingua scritta, ciò non accade; dunque lo «strale» si può, ma non si vuole trattenere. La percezione di cosa sia realmente *errore* è dunque diversa dalla nozione abitualmente impartita nella tradizione scolastica che ha, soprattutto in passato, stigmatizzato con particolare zelo gli errori ortografici, che invece qui sopravvivono in una (almeno) apparente noncuranza della norma.

È utile ricordare ancora che non siamo di fronte alle innumerevoli sviste legate alla velocità di scrittura ed alle sommarie riletture che abitualmente accompagnano l'interazione in rete[2], volte a non interrompere la fluidità della comunicazione[3]. Viceversa, proprio in direzione della velocità di scrittura, sarebbe lecito pensare ad una comprensibile necessità dell'utente di non perdere tempo in verifiche: il controllo dell'ortografia richiederebbe la consultazione di dizionari, enciclopedie o motori di ricerca. In una manciata di occorrenze (cfr. ess. [3a-c]) tale "pigrizia" è platealmente ammessa, senza riserve, persino con spiegazioni relativamente lunghe (quantomeno in relazione alla preponderante influenza della velocità di scrittura). È in realtà più importante non interrompere il flusso della scrittura: il gioco linguistico può poi dilatarsi senza rigide restrizioni.

[3a] Che bisogno c ' era di stravolgere la storia e di portare Frodo e Sam a osgillath (ora mi scoccio di controllare **come si scrive**)??? NUNC-IT Generic I.

[3b] Kwisatch Adecoso ... insomma quello li non saro' mai capace di ricordarmi **come si scrive** e sono troppo pigro per alzarmi e andare a cercare il libro . NUNC-IT Generic I.

[3c] Poi si scivolava direttamente , udite udite nella fine dell' ottocento (tra parentesi in particolare aborrisco Sostacovic (e non mi interessa di peritarmi di controllare **come si scriva**) e parte di quella musica sperimentale (o almeno io la chiamo cosi ') che caratterizza l' inizio Novecento la considero non musica bensi' sano , genuino - reditizio inquinamento acustico .) NUNC-IT Generic I.

Le formule esplicite "si scrive così" o "come si scrive", in forma di interrogative dirette od indirette, sembrano disimpegnare ed assolvere lo scrivente sia dall'errore, sia dalla noncuranza nei confronti di possibili fonti di controllo, quasi che la dichiarazione di consapevolezza dell'errore basti a giustificarli.

2.1 L'INTERROGAZIONE DEI NUNC. La ricerca per lemmi, impostata intorno a variazioni della query:

[4a] [lemma='scrivere'][word='così']

ha evidenziato contesti d'uso in cui i parlanti italiani sentono il bisogno di distanziarsi da ciò che hanno scritto, od almeno dalla forma grafica in cui l'hanno scritto, come nei seguenti esempi:

[4b] Cmq , bando alle ciancie (**si scrive così ?**) e cominciamo con il solito elenco di roba in uscita per Aprile ! NUNC-IT Generic II.

[4c] Con queste cose si fa ricerca , divertimento , passione , ma non professione , a meno che lo strumento per vendere sia " epatè le borsgiuà " (misero trucco per non dover scoprire **come si scrive**) NUNC-IT Generic I.

[4d] Beh , al punto seguente te lo dimostri io che " centra " (la **scrivo così** , alla Wess ..)) la Bor sa C' ENTRA con la politica . Eccome se c' entra !! NUNC-IT Generic II.

[2] Gli esempi che seguono, tratti interamente dai NUNC, riportano peraltro fedelmente (tokenizzazione a parte) eventuali refusi grafici all'interno dei post (norma che vale ovviamente per l'intera miscellanea, senza alcuna interpolazione successiva degli autori all'autenticità dei dati raccolti nei corpora), riproponendo anche errori di matrice evidentemente differente da quelli che ci interessano nel presente contributo.
[3] Cfr. Peticca 2002, cit. anche in Gheno 2005.

La ricerca ha in un primo tempo restituito diverse tipologie di referenti a cui rimanda "così". In un congruo numero di casi l'avverbio si riferisce al contenuto proposizionale, con usi cataforici:

[5] `Io ho` **`scritto così`** `: Spett. Rai , Facendomi interprete anche dei miei familiari e di molti conoscenti , chiedo che la manifestazione per la pace che si svolgerà a Roma sabato venga trasmessa in diretta .` NUNC-IT Generic II.

Da qui la necessità di esaminare in modo mirato le forme precedute dal "si" (query [6a]) e, in seconda battuta, l'interrogativa introdotta da "come" (query [6b]),

[6a] `[word='si'][lemma='scrivere'][word='così'],`
[6b] `[word='come'][word='si'][lemma='scrivere'],`

isolando i casi di mero dubbio ortografico che ci interessano in questa sede, per i quali comunque si è dovuto procedere ad una scrematura manuale dagli sporadici casi in cui l'interrogativa, diretta od indiretta, presentava uno *scope* più ampio:

[7a] `Andiamo per ordine : . distinzione netta : CROATO in caratteri latini , SERBO in caratteri cirillici . ^ difficoltà :` **`come si scrive`** `in Bosnia ?` NUNC-IT Generic I.
[7b] `[...] in un forum bisogna sempre stare attenti su` **`come si scrive`** `perchè non sappiamo come possano interpretare gli altri le frasi ..` NUNC-IT Generic I.

L'occorrenza più comune è l'interrogativa al presente indicativo, che coesiste con congiuntivi atti a rimarcare il dubbio, *che si scriva così?*, ed in frase subordinata, *mi chiedevo come si scrivesse*; non mancano poi esempi di futuro epistemico, *come si scriverà?*[4].

È utile notare a livello metodologico una inequivocabile diversificazione per varietà testuali nel numero di occorrenze. In newsgroup specialistici come quelli presenti in NUNC-IT Foto o Motori (ma compresi anche nei Generici, cfr. [8b]) gli esempi vertono su incertezze, se vogliamo, più plausibili, poiché relative a termini tecnici di basso uso al di fuori di contesti settoriali:

[8a] `Bene , l' effetto che vuoi tu è  ottenuto con un filtro fotografico che si chiama (ma che vergogna ! Non so nemmeno se` **`si scrive così`** `, ma penso di si ... ) cross-screen .` NUNC-IT Photo.
[8b] `Qualcuno può dirmi gentilmente cosè il filtro anisotropico ? A cosa serve ( per esempio l' antialaising , o` **`come capperi si scrive`** `, serve per " arrotondare " le forme dell' imagine di un videogioco )` NUNC-IT Generic I.

Più interessanti per i nostri scopi sono i newsgroup di carattere generale (raccolti in NUNC-IT Generic I e II), che non presuppongono tematiche necessariamente specialistiche (pur comprensibilmente presenti) ed il cui raggio d'azione si applica a contesti e co-testi spesso di linguaggio quotidiano. Emerge infatti una considerevole quantità di termini che nel parlato non determinerebbero alcuna perplessità, poiché la pronuncia ne è nota e generalmente riprodotta con sicurezza; si padroneggia il significato, ma è il significante a creare problemi di trasposizione grafica.

[4] Questo è evidentemente il motivo per cui ci siamo avvalsi nella query dell'interrogazione per `lemma` e non solo per `word`.

2.1.1 SI SCRIVE E SI ESITA. Le forme “si scrive così?”, “come si scrive?” e simili assolvono ad una funzione di distanziamento e forse di richiesta di conferma. Compare talvolta anche la parentetica “o come si scrive”, che mostra una sostanziale indifferenza alla soluzione ed all’atto stesso di commettere l’eventuale errore, pur ammettendo egualmente – dato altrettanto interessante a livello di coscienza metalinguistica del parlante – il dubbio ortografico:

[9] `Cmq non ho scritto perchè si è rotta la scheda madre ( mobo ) ma munendomi di tanta volontà di tanto scock ( o` **`come si scrive`** `) e di tanto carta stagnola ( sono un cuoco ) ho riparato la mobo . Ora ho il pc sta in una forma disumana sembra più la navicella di star treck che un pc .` NUNC-IT Generic I.

Non sono dell’avviso che gli utenti desiderino davvero essere rassicurati sulla grafia dagli altri partecipanti quando usano “si scrive così?” o simili note di riflessione linguistica. L’interrogativa parentetica sembrerebbe piuttosto un segno di presa di distanza, di *non-commitment* ed una concreta spia di quanto Simone definisce «enfatizzazione della fase processuale nel testo digitale» (Simone 2001, p. 45). Questo è vero per la pianificazione del testo, ma può valere anche per la grafia delle parole, come dimostrato dall’uso di formule di distanziamento come quelle analizzate.

Chi scrive osa forme che in un altro tipo di contesto produttivo non azzarderebbe proporre. L’ardimento può approdare al *divertissement*:

[10a] `ma sempre sta scia di sign devo cancellare .. Non ce le ho messe io , tutta colpa di quello che usava il computer prima di me , anzi il compiuter ,` **`si scrive così`** `, no ? ies , is ollrait !` NUNC-IT Generic II,

[10b] `" lo studente deve imparare aFILOSOFARE tramite la lettura dei testi dei grandi classici della filosofia ". Tipo quella lenza di Hegel ! O quel grande paraculo che è stato Nietzche ! ( oddio ,` **`come si scrive`** `? nice , non paraculo !) Ma se non sapete neppure cos' è il tempo e cosa significa esistere ed essere !` NUNC-IT Generic I.

Trattasi di una tendenza ordinaria nei contesti analizzati: talvolta per suscitare la risata degli altri partecipanti, certamente comunque per accattivarsene la simpatia, mostrando una competenza che si avvale in alcuni casi anche di regole morfologiche di formazione delle parole:

[11] `Lo scopo degli articoli e quello di sensibilizzare i ragazzini sul tema dell' abbandono e far venire voglia ai più grandicelli di fare volontariato . 1 Visto il target direi quelle dove si vedono giovani ragazze che si danno da fare per i cani , insieme a quelle dove i cani sono più peloucheosi (` **`si scriverà così`** `? ) .` NUNC-IT Photo Uncut.

Propenderei ad intravedere un risvolto ludico anche nei casi di incertezza sulla grafia dialettale, notoriamente soggetta a discussioni:

[12a] `Ricordatevi che accà nisciuno è fesso ( o` **`come`** `cacchio` **`si scrive`** `[...]` NUNC-IT Generic II,

[12b] `rumpiti i corn ! ( è calabrese non so` **`come si scrive`** `ma rende sempre l'idea )` NUNC-IT Generic I.

Talora traspare tuttavia un certo imbarazzo causato dall’insicurezza linguistica, accompagnato da esplicite scuse e scusanti,

[13a] `Ognuno ha una scheda di rete / funzionante e tutti sono in rete con uno switch / a porte ( non ricordo la marca ) . Da poco , il cliente ha fatto installare il lag ( non so se` **`si scrive così ,`** `chiedo venia qualora fosse sbagliato ) di Fastweb , che ha porte ethernet .` NUNC-IT Generic I,

[13b] `Anche questa catastrofe aveva la sua profezia , quella dei visionari di Mejugorie ( confesso , non so` **`come si scrive`** `)` NUNC-IT Generic I,

[13c] `Appena ho visto il trailer (` **`si scrive così`** `? comincio ad accusare il sonno ..) mi sono detta: " questo mai ".` NUNC-IT Generic I,

[13d] `Guarda forse soffri di haltzeimr (` **`si scrive così`**`? figuraccia ...): hai proprio detto così e mi spiace di aver cancellato quei post !` NUNC-IT Generic I,

od inviti alla verifica [13e], addirittura anteponendo [13f] l'interrogativa al termine dubbio, che è caso più marcato e meno frequente, avvertito ancor più quale *excusatio non petita*,

[13e] `[...] ( metto i nomi e i cognomi non per presunzione o pallosità ma perchè i vari Torquemada , sicuramente non` **`si scrive così ...`** `ma è lo stesso , possano risalire alla fonte e verificare ) [...]` NUNC-IT Generic I,

[13f] `beh , l ' ultimo gruppo citato è quello di un gruppo famosissimo di cui non mi ricordo` **`come si scrive`** `il nome .. tipo Mistique ma so gia' che non si scrive cosi' ! Mi potete dare un aiutino .. scusate l ' ignoranza .`
`Si chiamano Mis Teeq , e la lor o canzone di sottofondo è Scandalous` NUNC-IT Generic I.

La presenza di una difesa quasi "psicologica" sottolinea che in realtà il dubbio sulla resa grafica non è percepito con totale indifferenza, anche laddove questa sia spavaldamente ostentata dal turpiloquio o da altri intercalari inseriti nella stessa costruzione "come si scrive"; la query [14a] ne ha isolato in modo mirato varie casistiche, insieme a numerose, meno castigate[5], varianti, su cui sorvoliamo:

[14a] `[word='come'][pos='NOM'][lemma='scrivere']`

[14b] `Cioè dai sul serio vuoi metterti i led lampeggianti sul body ??? E' una delle cose più kitch , o` **`come diavolo si scrive`** `, che possa esserci ...` NUNC-IT Generic I.

[14c] `Per fare il mio divx ho utilizato clad dvd XP X esportare il dvd su hd FlasKMPEG XiS e il codec div x sempre passate e il codec mp della fraunaufer o` **`come cavolo si scrive`** `, non ricordo` NUNC-IT Generic I.

[14d] `Le scenografie . Buone , anche se ogni tanto , vuoi l' illuminazione " fredda " , vuoi le inquadrature particolari , nella testa sentivo risuonare un certo valzer ... ( ma` **`come cacchio si scrive`** `?!? )` NUNC-IT Generic I

[14e] `Devo cercare nei palinsesti il programma di economia di Alan Fridman e non so` **`come cippa si scrive`** `il nome di sto babbeo .` NUNC-IT Generic II.

[14f] `Sapete , c' ho un pò da fare al momneto , il Carnevale si avvicina =D`
`Ti vesti da mitocondro ( o` **`come caspita si scrive`** `) quest' anno ?` NUNC-IT Generic II.

---

[5] Su disfemismi e coprolalia nei newsgroup cfr. Gheno 2004, pp. 291-293. Sulla scarsità di eufemismi e forme di autocensura in rete cfr. anche Scholz 2003, p. 134.

Anche qui, in ultima analisi, ritroviamo un meccanismo di minimizzazione e ridimensionamento del dubbio. Non mancano comunque esempi in cui lo scrivente si imbatte in “maestrine dalla penna rossa” (e relativi dibattiti), con uno sguardo evidentemente molto lucido in [15c] ai meccanismi di riproduzione linguistica digitale:

[15a] Cazzo vuol dire free lands ? Terre libere ?
Se non sai **come si scrive** in inglese chiamali liberi professionisti o collaboratori sennò , fai uno sforzo , ti apri il dizionario di inglese e cerchi free-lance , che è il termine esatto ! Parlare male vuol dire pensare male !!!
NUNC-IT Generic I,

[15b] Sia ben chiaro che non sto qui montando in cattedra : io lessi solo il " Principe " anni fa , costretto da una cerbera di prof al liceo (scuola rigorosamente pubblica) , ma almeno questo mi insegno' a scrivere giusto il suo nome : ** Machiavelli ** . [...] Quindi , caro dotto , superbo , letterato , furbo ma non intelligente Andrea .... cos'è più importante ? Sapere **come si scrive** " Macchiavelli " o saper ragionare per penetrare nel SENSO DELLA VITA di cui tu non hai il minimo sentore ? P. S. Mi piacerebbe sapere cosa faresti tu , dopo anni fuori dall'Italia , in cui non senti più parlare italiano e che scrivi nella tua lingua madre solo rare volte . [...] Ma mi sorge il sospetto che tu utilizzi questa tattica della nozionistica ortografica perchè hai capito che è la cosa che mi fa inKAZ ... incavolare di più .... NUNC-IT Generic II,

[15c] * Trà * non **si scrive così** . Prova a vedere su se trovi informazioni riguardo l' italiano . Da notare che sulla tastiera del PC il tasto * a * ed il tasto * à * sono ben distanti tra loro , quindi il tuo non è stato un errore di battitura , sei proprio convinto che * trà * si scriva così .... NUNC-IT Generic I.

2.1.2 PAROLE “DIFFICILI”? Numerosi esempi di antroponimi attestano la strategia presa in esame di distanziamento e/o riflessione linguistica: si tratta di personaggi noti (attori, politici, personaggi celebri o di attualità) i cui nomi vengono graficamente alterati, nonostante occorrano ad altissima frequenza sulla carta stampata:

[16a] Peccato per il finale , che avrei sperato di vedere nella versione che rammentavo , veramente splendida , interpretata da Patrick Schwaize (**si scrive così** ?). NUNC-IT Generic II.

[16b] quello era un periodo strano , pieno di segreti e di paure ! anche lo scandalo mitrokin (**si scrive così**?) presentava scenari inqueietanti , [...] ! NUNC-IT Generic I.

[16c] Non solo de Filippi , ma anche Rai con la Deusanio (se **si scrive così**) . NUNC-IT Generic I.

Non si salvano dalle “storpiature” neanche i nomi di calciatori e sportivi in genere, malgrado il proverbiale (nonché effettivo) numero di quotidiani sportivi venduti in territorio italofono:

[16d] [...] quindi le finali di Capello furono incontri più equilibrati persi di misura , dove nella finale dell' Ajax guarda caso la differenza la fece un certo Reijkard (scusate se non **si scrive così**) , lo stesso giocatore che segnò a Monaco di Baviera . NUNC-IT Generic II.

Vero è che la difficoltà dello scrivente nella maggior parte dei casi restituiti dai NUNC (ess. [16a,b,d]) è legata a nomi stranieri, così come per alcuni toponimi:

[17] `Non era affar loro l'Italia . Era affar loro loro loro loro , lor loro , loro , casomai , il Giappone per la storia di Pearl Harbor ( o` **`come cavolo si scrive`** `) .` NUNC-IT Generic II.

Occorre pertanto distinguere due differenti gradi di difficoltà tra forestierismi e termini in lingua italiana, di cui ricorrono comunque numerosi esempi, che contemplano tra le altre alcune espressioni dell'italiano notoriamente ostiche e spesso soggette a grafie inesatte.

[18a] `Poi non so la tua mamma ma io mai li potrei cuocere senza prima averli passati sul retro della vecchia gratuggia (` **`come si scrive`** `?) per averli tutti con le gobbette che permettono al sugo di aderire meglio .` NUNC-IT Cooking,

[18b] `Mi trovi d ' accordo con te .. ( d'accordo` **`si scrive così`** `o daccordo ?)` NUNC-IT Generic I,

[18c] `[...] ; ma non scamperò il peana di " street spirit ", ma spero che il DAT fosse già bellechefinito [ ma` **`come si scrive`** `? bell'è che finito ? bel leche finito ?] -` NUNC-IT Generic I,

[18d] `La prossima volta mi presento a lezione con la fotocopia della fotocopia della scannerizzazione ( ma esiste questa parola , se sì ,` **`si scrive così`** `?) della fotocopia della fotocopia dell'originale .` NUNC-IT Generic I.

Tra i forestierismi numerose occorrenze derivano *in primis* da inglese e francese, quali lingue statisticamente più frequenti nei prestiti nostrani (ma certamente anche per una maggiore distanza rispetto all'italiano tra sistema fonologico e sistema di scrittura[6]):

[19a] `Tu non contribuisci ad una sega su questo newsgrouop a parte sparare a zero su tutto e tutti , * newsgrouop * non` **`si scrive così`** `... ma dopotutto , è una parola inglese ...` NUNC-IT Generic I,

[19b] `no è che capita durante i zapping da dopo lavoro ,` **`si scrive così`** `o si scrive zupping ?` NUNC-IT Generic II,

[19c] `ciao a tutti avrei qualche domanda [.] per estrarre il negativo dal rullo si usa lo stappabottiglie ( non il tirabouchot o` **`come si scrive`** `), quale tappo fate saltare ?` NUNC-IT Photo,

[19d] `Qui ti potrei dare un minimo ragione dicendo che l ' idea base è banale , ma secondo me è stata gestita molto bene , sopratutto con l ' escamotage ( so che non` **`si scrive così`** `ma sono ignorante ! ) della proprietaria del ristorante [...]` NUNC-IT Generic II,

e non mancano anche meno frequenti ma altrettanto problematici termini da lingue tipologicamente distanti,

[20a] `Allora attenzione a come addestrate il chiuaua ( o` **`come si scrive`** `)` NUNC-IT Generic I.

[20b] `Poi fu abbandonato l ' idea del socialismo , e c'è stato un nuovo rafforzamento della Shiad ( nn so` **`come si scrive`** `) Islamica .` NUNC-IT Generic I.

---

[6] Cfr. anche, tra le altre, le riflessioni sulla fonologia (p. 160) nel recente intervento di Scalise - Ceccagno 2006 sulla possibilità di definizione di lingue “facili” e “difficili”.

per tacer, naturalmente, dei termini in lingua russa od altre lingue slave, comunque presenti tra i risultati estratti, che pongono allo scrivente italiano l'ulteriore controversa questione della traslitterazione dai caratteri cirillici[7].

Alcuni dubbi, poi, provengono dal latino,

[21a] `L ' argomento pace è super partes ( o` **`come si scrive`** `) , quindi mi sembra il caso di postarlo ovunque .` NUNC-IT Generic II.

[21b] `La schiavitu e la libertà sono una forma mentis ( o` **`come`** `cavolo` **`si scrive`** `) , in questo caso non c'è nessun vincolo se non nel proprio " io ".` NUNC-IT Generic II.

e riscontriamo, inoltre, una serie di termini stranieri arbitrariamente italianizzati nella grafia,

[22a] `Dopo la corsa sul tapirulan (` **`si scrive così`** `?) e la lezione di GAG ( gambe addominali e glutei ) stavo pedalando alla cyclette sudata e stanca più che mai .` NUNC-IT Generic II,

[22b] `Pregi dell' arrivo del Pendolino ? pagato poco , stipendio ( relativamente ) basso , lascia spazio ad altri arrivi ben più onerosi . vero . il soprannome fa pandan ( o` **`come`** `ca o` **`si scrive`** `) con il Concorde dall' altra parte .` NUNC-IT Generic II,

[22c] `sono arrivato in silenzio e parlando sottovoce , non voglio certo venire alla ribalta per aver scatenato flame con gente che nemmeno è un abituè e (` **`si scrive così`** `?) di ISCR . torno nei ranghi .` NUNC-IT Generic II,

od adattati, per esempio con conseguente coniugazione verbale[8],

[22d] `è circa un anno che non scrivo su questo newsg, più che altro lurkavo (`**`si scrive così`**`?) ...` NUNC-IT Generic II.

Curiosi, infine, alcuni casi di errata segmentazione come quello in [18c] ed il seguente:

[23] `La vicende è quella di un'ossessione , la ricerca spasmodica di un ordine matematico in ogni aspetto della vita , che parte dal controllo della borssa per poi finire all' Atohra ( o` **`come si scrive`** `) e a Dio stesso .` NUNC-IT Generic I.

2.2 IL SIGNIFICANTE TRA LIVELLO FONETICO E LIVELLO GRAFICO. Il sistema di interrogazione di cui sono corredati i NUNC ha permesso anche l'estrazione di termini intervallati da un numero definito di parole; grazie ai risultati della query

[24] `[word='si'][]{0,1}[word='così']`

è stato possibile distinguere tra "scrivere" da un lato e "dire" o "chiamare" dall'altro, constatando (malgrado alcune ovvie sovrapposizioni di contesti) un effettivo discrimine tra le occorrenze in cui la parola appropriata è nota, e problematica è piuttosto la traduzione grafica della rappresentazione sonora (cfr. esempi in § 2.1.1 e 2.1.2 con il verbo "scrivere"), e quei contesti in cui invece manca completamente la conoscenza del termine opportuno (ignoranza di "come si dice", "come si chiama" negli esempi che seguono, e non ignoranza della resa grafica).

---

[7] Giusto un piccolo esempio, a mo' di assaggio:

[20c] `Quando Fester e Gomez duellano in sala pranzo , la musica che li accompagna è una Danza Ungherese di Brahms o la " Danza delle Spade " di Kachturian ( o` **`come cavolo si scrive`** `) ??` NUNC-IT Generic I.

[8] Nonché derivabilità morfologica: *lurkaggio*, *lurkatore*; su *lurkare* e simili adattamenti, cfr. Valle ¶ 16, *infra*.

[25a] Ora appena mi sgessano volevo tornare a scivolare sulla neve (senza uccidermi però) , ma questa volta con una protezione in più : un bel paio di guanti con i parapolsi (**si dice così** ??). NUNC-IT Generic I.

[25b] E poichè i contratti d ' affitto non sono in genere annuali ma quadriennali o ottennali (**si dice così** ?), se la media è [...] l' incremento sarà molto più alto . NUNC-IT Generic I.

[25c] rete interna di Pc (" server " win e clients win)collegati con cavo bnc (credo **si chiamino così**) , non col cavetto tipo telefonico , per intenderci meglio . NUNC-IT Generic I.

Talvolta dubbio sul significato e dubbio sul significante si sovrappongono, fino a proporre agli interlocutori una spiegazione del senso:

[26a] il sapore e la consistenza delle MozartKugeln (**si scrive così** ? comunque sono i cioccolatini Mozart , quelli con il cioccolato fuori e il nucleo di marzapane) certo , quelli che riempiono le vetrine a Salisburgo e nn solo (hanno il pistacchio anche , no ?) NUNC-IT Generic I,

[26b] Ciao a tutti ho un problema con Win xp professional in pratica mi succede che il file explorer . exe (questo lo vedo dal task manager) mi va in " loop " (credo **si scriva così** cmq in ciclo senza fine) NUNC-IT Generic I.

Di séguito alcuni esempi in cui la consapevolezza dei parlanti rispetto al dislivello grafia-pronuncia è particolarmente marcata:

[27a] [...] e per finire ci hanno portato il Sachè (non so **come si scrive** ... io l ' ho scritto così come si pronuncia ! : è proprio amaro . NUNC-IT Generic I.

[27b] vorrei fare un regalo a una bimba di - anni ha avuto un bello spavento tempo fa e ha paura folle dei cani ovviamente non un cucciolo di (non so **come si scrive** e ve lo scrivo come si dice !) " rotvailler " ;) NUNC-IT Generic I.

Dagli esempi citati emerge la coscienza dello scrivente del diverso piano tra ortografia e pronuncia, e dunque di un passaggio alla realizzazione (orto)grafica[9] del materiale fonologico non necessariamente privo di asperità. L'utente ripropone quel salto dalla *suppositio formalis* a quella *materialis* che Conte 1999 definiva metalinguistico – riferendosi il parlante ad un type e non ad un token[10] – e mette in atto uno «shift in the domain of reference, from 'the world' to language», come richiamato da De Brabanter 2004[11].

Il discorso metalinguistico, seguendo Rettig 1976, pp. 61-63, può essere di due tipi: uno, tipico del linguaggio corrente, equivalente alle opinioni espresse sulla lingua dal parlante ingenuo; l'altro, tipico del linguaggio di grammatici e linguisti, quale comprensione di oggetti prevalentemente od esclusivamente linguistici. I due piani non sarebbero diversi nei fondamenti, la distinzione è formulata da Rettig in termini di attualità e grado di istituzionalizzazione :

---

[9] Realizzazione che può trovare, in qualche caso, anche difficoltà di ordine pratico, di digitazione sulla tastiera:

[27d] [...] uno dei cori di incitamento provenienti dalla allora Curva Filadelfia era : " Goba ", con dieresi sulla " o " che non so **come cavolo si scrive** . In piemontese significa " gobba " , da intendersi come vecchia , da intendersi come " vecchia signora (del calcio italiano)" . NUNC-IT Generic II.

[10] Cfr. il noto es. (Lyons 1977, I. p. 667): *(X says) That's a rhinoceros (and Y responds) A what? Spell it for me.*

[11] Che pure trascura, a mio avviso, risultati recenti considerevoli nel campo della linguistica testuale, come le soluzioni proposte dalla stessa Conte.

> *«Man kann sich einen stufenlosen Übergang von der leicht hingeworfenen Bemerkung in der alltäglichen Rede über das Aufnotieren von aphorismenhaften Betrachtungen, über die regelmäßig publizierte Sprachchronik in einer Zeitung bis hin zur Schulgrammatik und ausführlichen wissenschaftlichen Grammatik und zum Linguistenkongreß denken. Dennoch halten wir es für sinnvoll, zwischen diesen beiden Typen zu differenzieren, denn sie unterscheiden sich wesentlich in bezug auf Aktualität und Grad der Institutionalisierung.»* (Rettig 1976, p. 61)

L'espressione metalinguistica definita "quotidiana" è riconducibile ai casi rilevati nei newsgroup ed è, rispetto alla questione dell'attualità avanzata da Rettig, nata da un'occasione contingente; rispetto al secondo parametro, è, diversamente dal linguaggio scientifico-linguistico, poco istituzionalizzata: proprio nel passaggio tra queste polarità possiamo intravedere una traccia di quella "coscienza metalinguistica" che abbiamo ipotizzato nell'utente di newsgroup, senza la quale il dubbio ortografico verrebbe semplicemente ignorato e non marcato da un'interrogativa.

3. WIE SCHREIBT MAN ES? Disponendo di corpora comparabili ai NUNC in lingua francese, inglese, spagnola e tedesca, sarà significativo indagare se le incertezze sulla grafia siano peculiari dei frequentatori di newsgroup in italiano o caratterizzino anche stranieri alle prese con dubbi simili, rispetto sia a termini della propria madrelingua sia a prestiti lessicali e calchi.

È possibile ipotizzare percorsi simili anche nelle altre lingue, visto che il fenomeno ci sembra legato ad un'istanza *language-independent* di (a) riflessione metalinguistica e (b) passaggio dal piano fonologico al piano ortografico.

L'interrogazione, in particolare, del NUNC in lingua tedesca lascia intravedere interessanti orizzonti in relazione alla riforma ortografica. Modificata più volte nel giro di pochi anni[12], la riforma ha contribuito ad alimentare in ambiente germanofono i dubbi dei parlanti nativi rispetto a diverse questioni: terreno in cui risulta ancor più evidente quanto l'insicurezza sia ortografica[13] piuttosto che semantica o morfologica – circoscritta in questo caso a termini della propria madrelingua e solo parzialmente legata ai forestierismi. Oltre quindi ad esempi molto simili ai contesti riscontrati nei newsgroup italiani, come [28], si trovano casi come [29]:

[28a] **Wie schreibt man** nochmal Tschai [phon.] ? NUNC-DE Generic,
[28b] Gedealt wird dort primär
mit Koks und Amphetaminen (**wie** auch immer **man** die inzwischen **schreibt** ;-)) - was gerne schon bei Durchsuchungen bei der Eingangskontrolle auffliegt . NUNC-DE Generic,
[28c] weil den Tod von Cedrick (weiss nicht meh genau **wie man** ihn **schreibt**) im Band 4 NUNC-DE Generic:

[29] Hab ich die Rechtschriebreform 2003 verpasst ? Zur meiner Schulzeit **hieß es** noch : wer nämlich mit " h " schreibt ist dämlich . ;) NUNC-DE Generic.

Eisenberg - Fuhrhop 2007, p. 25, vedono nell'errore di corrispondenza grafema-fonema ("GPK-Fehler" = *Graphem-Phonem-Korrespondenz-Fehler*) un prototipo dell'errore ortografico, deviazione dal principio di scrittura alfabetica. Problematica diventa allora la coincidenza della didattica dell'ortografia con la didattica della scrittura (*ibidem*, p. 18), tanto più che, in séguito alla riforma ortografica, si finisce a scrivere «systemwidrig, obwohl er orthographisch korrekt schreibt»[14]: gli errori contro la *Norm* hanno valenza diversa dagli errori di *System*.

[12] Le norme dell'ultima *Rechtschreibreform* del 1996 sono state riviste nel 2004 e nel 2006.
[13] Si noti come una nota rivista di linguistica, la "Zeitschrift für Sprachwissenschaft", dedichi nel 2007 un intero *Jubiläumsheft* al problema dell'ortografia tedesca: chiaro sintomo di una tematica tuttora complessa e delicata.
[14] Cfr. l'esempio riportato a p. 26: se un bambino producesse l'enunciato "Der Strom brummt mit fünfzig Herz" con *Herz* invece di *Hertz*, ciò non sarebbe considerato da Eisenberg - Fuhrhop *Systemfehler*, bensì come una di quelle deviazioni dalla norma su cui è maggiormente intervenuta la riforma ortografica.

4. SCRITTURA E GRAMMATICA NORMATIVA. La priorità ontologica del parlato rispetto allo scritto spiega certamente parte della questione esaminata, ma sono assai indicativi anche post come il seguente:

[30] Dimostri una profonda ignoranza chiamandoci fiorenzia , si dice Florentia , e non dirmi che non sai il latino perchè anche se così fosse dovresti saper leggere alla tua età .
su questo hai ragione scusa ! ma di solito leggo la gazza fino a pagina - , le notizie (ops , scusa , trafiletti scritti in minuscolo un giorno si e due no) sulla fiorenza me le sono perse ! onestamente non ho mai visto **come si scrive ,** considerando anche il fatto che avete cambiato nome una decina di volte . NUNC-IT Generic I,

che suggerisce l’esistenza di un deposito di termini immagazzinati dal parlante a livello acustico e non visivo: parole sentite ma non lette.

Se la fruizione delle informazioni passasse unicamente da canali acustici quali televisione, radio e telefono (certamente rilevanti nella comunicazione), la difficile trasposizione grafica sarebbe comprensibile, ma i crescenti indici di utilizzo di Internet dovrebbero invece inficiare questa ipotesi: la rete mette a disposizione del madrelingua italiano una quantità enorme di testi in forma scritta (e molto più raramente orale).

Simone 2000 vede nella scrittura digitale online una “terza fase”, uno stadio in cui l’intelligenza non lavorerebbe più in modo *sequenziale* come nella lettura, ma si starebbe abituando a procedimenti olistici processati da un’intelligenza *simultanea*, in grado di gestire contemporaneamente più informazioni, senza però essere in grado di stabilire in essi una gerarchia od un ordine. La dimostrazione di una metamorfosi cognitiva come quella delineata da Simone è, a mio avviso, assai ardua; l’idea di una tendenza della società odierna a *sentire* più che a *vedere*, come alcuni mezzi di comunicazione di massa hanno stimolato a fare, farebbe retrocedere la vista dalla decodificazione di segni grafici alla fruizione meno consapevole di immagini e movimenti sul monitor, in modo meno analitico di quanto richiesto dalla lettura lineare di un intero libro. Uno scritto digitale fatto di testi brevi ed immagini ed accompagnato da una ricerca interattiva di informazioni che si sostituisce spesso alla lettura meditata potrebbe fornire una prima ipotesi esplicativa del così elevato numero di grafie scorrette.

Alcuni interessanti studi psicolinguistici, tuttavia, in particolare in seno alla ricerca sul sistema di scrittura giapponese e sulla processabilità cognitiva dei *kanji* (logogrammi) e dei *kana*, hanno mostrato che la lettura non avviene alfabeticamente: si tratta piuttosto di una decodifica “logografica” del segno scritto (in cui peraltro non è ancora chiaro, così come per gli *hanzi* cinesi, se l’informazione fonologica sia interpretata ad un livello pre- o post-lessicale; cfr. Kess - Miyamoto 1999, pp. 34-57), ma per la quale conterebbe il segno linguistico *in toto* prioritariamente alla nozione di grafema.

Se tali dati non sono necessariamente trasferibili in modo diretto a tutti i sistemi di scrittura, mi pare però che il discorso possa essere ricondotto agli argomenti efficacemente impostati già negli anni ’80 da Cardona a proposito della scrittura. Muovendo da un sano avvertimento a non definire la scrittura storicamente a ritroso (in modo alfabetocentrico, il che comporterebbe un’ottica solo “occidentale”, a scàpito delle culture che nei secoli hanno sviluppato sistemi grafici non alfabetici[15]), Cardona si addentra in un’interessante analisi antropologica del fenomeno *scrittura* e ricorda che «ogni società esprimerà quei tipi di scrittura che le saranno congeniali e necessari o ne adotterà di esterni» (1981, pp. 22-23). Superando la prima fase della linguistica moderna che considerava la scrittura semplicemente come «specchio più o meno fedele della lingua parlata […], come sequenza di segni che trascrivono suoni della lingua» (*ibidem*, p. 19),

[15] Come riproposto recentemente, seppur in altri termini, anche da Diamond 1999 (§ 12, pp. 215-238).

Cardona parte dall'idea di una materia di per sé amorfa, che viene segnata dalla lingua, dall'intenzione semiotica. Non si può dire che pensiero e lingua coincidano; piuttosto che la lingua lascia tracce nel pensiero. Non importa allora qual è il sistema semiotico con cui si vuole comunicare, il meccanismo ontologico resterà il medesimo, sia che si tratti di sistemi grafici cuneiformi, sia di scrittura digitale online. In questa prospettiva la grammatica normativa, per quanto indispensabile, assume una valenza diversa.

Ciò che tuttavia resta maggiormente da notare è relativo ad una distinzione sostanziale tra parlare e scrivere: i parlanti pensano e parlano allo stesso tempo; lo scrivente dovrebbe prima pensare e in una fase successiva mettere per iscritto. In questo punto si colloca la discriminante dell'utente newsgroup: tendenzialmente, nel flusso del discorso, egli non distingue più le due fasi e pensa scrivendo, riducendo la pianificazione del discorso ed offrendoci un importante elemento, finora poco considerato nella letteratura, per parlare davvero di "vicinanza alla lingua orale".

5. CONCLUSIONI. Non esiste un criterio oggettivo ed univoco per stabilire quali siano le parole "difficili" da scrivere, a parte alcune consuetudini didattiche adottate nella scuola dell'obbligo e cristallizzate nel sentire comune, dato il peso notevole affidato alla correttezza ortografica in tutte le esperienze scolastiche di letto-scrittura.

Valicare il confine della semplice oralità e trasporre graficamente l'informazione linguistica implica operazioni cognitive diverse, di cui si è qui voluta analizzare una difficoltà specifica, legata ad uno solo dei tre vertici del triangolo semiotico: referente e significato non pongono problemi nei contesti esaminati, il piede malfermo del triangolo è costituito dal significante.

La veste grafica con cui questo si presenta è sintomo interessante di una lingua in evoluzione, anche sul piano normalmente più refrattario alla trasformazione, quello della lingua scritta, che in rete riscontra una più energica infiltrazione ed accettazione dell'errore "classico".

Questi primi appunti sull'argomento mostrano la preminente utilità dei NUNC per lo studio della lingua in rete e per l'analisi dell'interazione tra norma e sistema nella lingua italiana standard e neostandard.

BIBLIOGRAFIA.

ALLORA

2003 Adriano Allora, *È scritto o parlato?,* in "Italiano & Oltre" I (2003) 14-18.

2005 Adriano Allora, *A Tentative Typology of Net Mediated Communication*, comunicazione presentata alla *Corpus Linguistics 2005 Conference, Birmingham July 14-17 2005*, disponibile online alla pagina `http://www.corpus.bham.ac.uk/PCLC/`

ALLORA - MARELLO

i.p. Adriano Allora - Carla Marello, *"Ricarica clima". Accorciamenti nella lingua dei newsgroup*, contributo per il *IX congresso internazionale della Società di linguistica e filologia italiana (SILFI). Prospettive nello studio del lessico italiano, Firenze 14-17 giugno 2006*, in corso di stampa negli *Atti*.

ANDORNO

2003 Cecilia Andorno, *Linguistica testuale. Un'introduzione,* Carocci, Roma, 2003 "Università" 519.

BARBERA

2007 *i.s.* Manuel Barbera, *I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo*, in "Cuadernos de filología italiana" XIV (2007) 11-32, in corso di stampa.

¶ 1 Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in questo volume, pp. 3-20.

BERRUTO

1985 Gaetano Berruto, *Per una caratterizzazione del parlato: l'italiano parlato ha un'altra grammatica?*, in HOLTUS - RADTKE 1985, pp. 120-153.

2005 Gaetano Berruto, *Italiano parlato e comunicazione mediata dal computer*, in HOELKER - MAAß 2005, pp. 109-124.

BOSC - MARELLO - MOSCA

2006 *Saperi per insegnare. Formare insegnanti di italiano per stranieri. Un'esperienza di collaborazione tra università e scuola*, a cura di Franca Bosc - Carla Marello - Silvana Mosca, Torino, Loescher 2006 "Università degli studi di Torino - Ufficio scolastico regionale per il Piemonte".

CAFFI

2001 Claudia Caffi, *La mitigazione. Un approccio pragmatico alla comunicazione nei contesti terapeutici*, Münster, LIT, 2001.

CARDONA

1981 Giorgio Raimondo Cardona, *Antropologia della scrittura*, Torino, Loescher, 1981.

CONTE

1999/88 Maria-Elisabeth Conte, *Condizioni di coerenza*, Alessandria, Edizioni dell'Orso, 1999. Nuova edizione, con l'aggiunta di due saggi a cura di Bice Mortara Garavelli, di Maria-Elisabeth Conte, *Condizioni di coerenza. Ricerche di linguistica testuale*, Firenze, La Nuova Italia Editrice, 1988 "Pubblicazioni della Facoltà di Lettere e filosofia dell'Università di Pavia" 46.

CORINO

¶ 13 Elisa Corino, *NUNC est disputandum. Aspetti della testualità e questioni metodologiche*, in questo volume, pp. 225-252.

COVINO

2001 *La scrittura professionale*, a cura di Sandra Covino, Olschki editore, Firenze, 2001.

DE BRABANTER

2004 Philippe De Brabanter, *'World-to-Language' Shifts between an Antecedent and its Pro-Form,* preprint, disponibile online alla pagina `http://jeannicod.ccsd.cnrs.fr/docs/00/05/36/07/PDF/ijn_00000555_00.pdf`

DIAMOND

1999 Jared Diamond, *Guns, Germs, and Steel: the Fates of Human Societies*, New York - London, W. W. Norton & Company, 1999.

DIETER

2006 Jörg Dieter, *Webliteralität. Lesen und Schreiben im World Wide Web*, Inauguraldissertation zur Erlangung des Grades eines Doktors der Philosophie im Fachbereich Neuere Philologien (10) der Johann Wolfgang Goethe-Universität zu Frankfurt am Main, Einreichungsjahr: 2006, Erscheinungsjahr: 2007, disponibile online alla pagina `http://www.webrhetorik.de/Arbeit/arbeit.html#Download`.

EISENBERG - FUHRHOP

2007 Peter Eisenberg - Nanna Fuhrhop, *Schulortographie und Graphematik*, in "Zeitschrift für Sprachwissenschaft", Jubiläumsheft XXVI (2007) 15-41.

FIORE - MCLUHAN → MCLUHAN - FIORE.

FIORENTINO
2004 Giuliana Fiorentino, *Scrivere come si parla - Variabilità diamesica e CMC: il caso dell'*e-mail, in "Horizonte" 8, 83-110.
2005 Giuliana Fiorentino, *Così lontano, così vicino: coerenza e coesione testuale nella scrittura in rete***,** in KORZEN 2005, solo nel Cd-rom allegato.

GHENO
2004 Vera Gheno, *Prime osservazioni sulla grammatica dei gruppi di discussione telematici di lingua italiana*, in "Studi di Grammatica Italiana" XXII (2004) 267-308.
2005 Vera Gheno, *Alcune "metamorfosi" linguistiche nei gruppi di discussione telematica*, in «www.scriptamanent.net», anno III (2005)$_{21\text{-giugno}}$, disponibile online alla pagina `http://www.scriptamanent.net/scripta/public/dettaglioNewsRivista.jsp?ID=1000921`

GOFFMANN
1964 Erving Goffmann, *The Neglected Situation*, in "American Anthropologist" VI (1964)^{2} 133-136.

GÜNTHER - LUDWIG
1994 *Schrift und Schriftlichkeit | Writing and Its Use*, herausgegeben von | edited by Hartmut Günther und | and Otto Ludwig, Berlin, de Gruyter, 1994.

HERRING
1996 *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives* edited by Susan C. Herring, Amsterdam, John Benjamins, 1996.

HOELKER - MAAß
2005 *Aspetti dell'italiano parlato. Tra lingua nazionale e varietà regionali. Atti del congresso tenuto ad Hannover, 12-13 maggio 2003*, a cura di Klaus Hoelker - Christiane Maaß, Münster - Hamburg - London, LIT-Verlag, 2005 "Romanistische Linguistik" 6.

HOLTUS - RADTKE
1985 *Gesprochenes Italienisch in Geschichte und Gegenwart*, herausgegeben von Günther Holtus und Edgar Radtke, Tübingen, Narr, 1985 "Tübinger Beiträge zur Linguistik" 252.

KESS - MIYAMOTO
1999 Joseph F. Kess - Tadao Miyamoto, *The Japanese Mental Lexicon: Psycholinguistic Studies of Kana and Kanji Processing*, Philadelphia - Amsterdam, John Benjamins, 1999.

KOCH - ÖSTERREICHER
1985 Peter Koch - Wulf Österreicher, *Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte,* in "Romanistisches Jahrbuch" XXXVI (1985) 15-43.
1994 Peter Koch - Wulf Österreicher, *Schrift und Schriftlichkeit,* in GÜNTHER - LUDWIG 1994, pp. 587-604 (Artikel 44).

KORZEN
2005 *Lingua, cultura e intercultura: l'italiano e le altre lingue. Atti del VIII convegno Silfi, Società Internazionale di Linguistica e Filologia Italiana (Copenaghen, 22-26 giugno 2004)*, a cura di Iørn Korzen, Copenaghen, Samfundslitteratur Press, 2005 "Copenhagen Studies in Language" 31.

LYONS
1977 John Lyons, *Semantics*, 2 voll., Cambridge - London - New York - Melbourne, Cambridge University Press, 1977.

McLuhan
1964/90 [Herbert] Marshall McLuhan, *Understanding media*, New York, McGraw-Hill Book Company, 1964. Trad. it. *Gli strumenti del comunicare*, Milano, Il Saggiatore, 1990.

McLuhan - Fiore
1968 [Herbert] Marshall McLuhan - Quentin Fiore, *War and peace in the global village; an inventory of some of the current spastic situations that could be eliminated by more feedforward*, co-ordinated by Jerome Agel, New York, McGraw-Hill, 1968.

Ong
1982/86 Walter J. Ong, *Orality and Literacy. The Technologizing of the Word*, London - New York, Methuen, 1982. Trad. it. di Alessandra Calanchi, *Oralità e scrittura. Le tecnologie della parola*, Bologna, Il Mulino, 1986.

Peticca
2002 Sara Peticca, *Il linguaggio dell'email*, Soveria Mannelli, Rubbettino, 2002 "La politica".

Presch - Gloy
1976 *Sprachnormen II. Theoretische Begründungen - außerschulische Sprachnormepraxis*, herausgegeben von Gunter Presch und Klaus Gloy, Stuttgart - Bad Cannstatt, Friedrich Frommann Verlag - Gunther Holzboog KG, 1976.

Rainer - Stein
2003 *I nuovi media come strumenti per la ricerca linguistica*, herausgegeben von Franz Rainer und Achim Stein, Frankfurt am Main, Peter Lang, 2003.

Rettig
1976 Wolfgang Rettig, *Sprachnormen und Systemlinguistik*, in Presch - Gloy 1976, pp. 50-70.

Saussure
1916/67/95 Ferdinand de Saussure, *Cours de linguistique générale*, publié par Charles Bailly et Albert Séchehaye, avec la collaboration de Albert Riedingler, édition critique préparée par Tullio de Mauro, postface de Louis-Jean Calvet, Paris, Payot, 2001^{r} [1995_3, 1972_1] "Grande bibliothèque Payot". Edizione originaria: *ibidem*, 1916. Edizione italiana: *Corso di linguistica generale*, introduzione traduzione e commento di Tullio De Mauro, Roma - Bari, Laterza, 1967_1.

Scalise - Ceccagno
2006 Sergio Scalise - Antonella Ceccagno, *Facile o difficile? Alcune riflessioni su italiano e cinese,* in Bosc - Marello - Mosca 2006, pp. 153-177.

Scholz
2003 Arno Scholz, *Comunicazione giovanile in rete: una mailing list italiana dedicata alla cultura hip-hop*, in Rainer - Stein 2003, pp. 117-139.

Simone
2000 Raffaele Simone, *La terza fase. Forme di sapere che stiamo perdendo*, Bari, Laterza, 2000.

2001 Raffaele Simone, *Tre paradigmi di scrittura,* in Covino 2001.

Valle
¶ 16 Luca Valle, *Ricerche su anglismi nei NUNC francesi e italiani. Tra "lurker", "lurkeur" ed altri prestiti*, in questo volume, pp. 285-296.

CORPORA DI RIFERIMENTO[16].

corpora.unito.it	`http://www.corpora.unito.it/`
NUNC	`http://www.bmanuel.org/projects/ng-HOME.html`

[16] Aggiornati al 28 febbraio 2007.

CRISTINA ONESTI - MARIO SQUARTINI[1]
(Università di Torino)

15. "Tutta una serie di"[2]. *Lo studio di un pattern sintagmatico e del suo statuto grammaticale.*

0. PREMESSA. L'individuazione di combinazioni sintagmatiche ricorrenti ha tradizionalmente rappresentato un banco di prova privilegiato per quel settore della linguistica dei corpora in cui la più tradizionale ricerca lessicografica su combinazioni idiomatiche e fraseologia si è fusa con la riflessione teorica sulla definizione nozionale di lessema polirematico o di collocazione[3]. D'altra parte, accanto allo studio dei fenomeni di più diretta rilevanza lessicografica, un filone parallelo della linguistica dei corpora si è concentrato su strutture sintagmatiche di pertinenza più propriamente "grammaticografica", in cui cioè la collocazione non riguarda la coricorrenza di singole entrate ma un pattern strutturale ricorrente. Ancora in anni fondativi per la linguistica dei corpora Renouf - Sinclair 1991 hanno significativamente dedicato un lavoro ai "collocational frameworks", cioè a quelle collocazioni sintagmatiche che si basano su un modello strutturale replicato da diverse entrate lessicali, ad esempio il sintagma nominale inglese "*a* + __ + *of*" in cui il segnaposto può essere riempito da una lista di lessemi (*couple*, *series*, *pair*, *lot*, *piece*, *quarter*, *variety*, *member*, *number*, *kind*, *sort*, *matter*, *result*, ecc.) che non formano una classe naturale dal punto di vista semantico (mentre alcuni di essi quantificano, ad esempio *couple*, *number*, altri "qualificano" il nome con cui si collocano, ad esempio *kind*, *sort*). La natura autenticamente *corpus-driven* dello studio di questo tipo di pattern collocazionale è evidente ed il loro ruolo in un modello di lessico-grammatica delle costruzioni ("pattern grammar", "construction grammar") è stato ampiamente dimostrato (si veda ad esempio Hunston - Francis 2000).

In questo lavoro partiamo dal presupposto che lo studio di schemi collocazionali di questo genere, al di là delle prospettive generali per una definizione olisticamente "costruzionalista" della grammatica, sia più modestamente applicabile come prerequisito per l'analisi di singoli fenomeni grammaticali, in particolare per quei fenomeni il cui statuto categoriale risulti per qualche motivo controverso o la cui rilevanza "grammaticografica" sia passata inosservata. In questa prospettiva abbiamo scelto un possibile pattern italiano che rappresenta un sottoinsieme di quello studiato da Renouf - Sinclair 1991 per l'inglese ("*a* + __ + *of*"), riducendone così la portata euristica ma allo stesso tempo rendendo più stringente la possibilità che esso individui una classe naturale di oggetti di rilevanza (lessico-)grammaticale. Rispetto a Renouf - Sinclair 1991 il pattern è stato quindi ulteriormente specificato nelle sue restrizioni collocazionali aggiungendo anche un modificatore aggettivale nella forma seguente: "tutto/a + un(a) + __ + di".

Come argomentiamo nel § 3 lo statuto categoriale di alcune delle sequenze individuate da questo pattern è stato variamente interpretato nella letteratura sull'argomento. Il nostro studio intende prima di tutto verificare se i lessemi che occupano la posizione non specificata, cioè il segnaposto del pattern, abbiano tratti comuni che permettano di individuare una corrispondenza non casuale tra sequenza sintagmatica e funzione grammaticale. In una seconda fase del lavoro

---

[1] I §§ 1, 2 e 4 sono da attribuire a Cristina Onesti, mentre Mario Squartini si è occupato della stesura dei §§ 0 e 3.
[2] Nell'àmbito della giornata di studi, la comunicazione di cui questo articolo è sostanziale rielaborazione era più laconicamente intitolata *"Tutta una serie di": esempi da corpora diversi.*
[3] Il ricco interscambio tra linguistica dei corpora e pratica lessicografica nell'àmbito della fraseologia è ben documentato dala sezione *Phraseology and Collocation* nel recente Corino - Marello - Onesti 2006, pp. 909-1087.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 271-284.

abbiamo approfondito le caratteristiche strutturali del pattern individuato cercando nei corpora dati significativi per dirimerne l'attribuzione ad una specifica categoria grammaticale (nomi, determinanti, quantificatori, classificatori?). Le due fasi del lavoro corrispondono anche a due procedure distinte: l'individuazione del pattern, che viene presentata nel § 2, risponde ad una metodologia più propriamente *corpus-driven*, mentre l'approfondimento sulla natura categoriale (§ 4) esemplifica un utilizzo più tradizionalmente *corpus-based* dei corpora, che sono stati interrogati rispetto alle caratteristiche morfosintattiche di una specifica realizzazione del pattern, quella che i risultati della prima fase della ricerca avevano permesso di individuare come la più frequente nei corpora italiani analizzati: *tutta una serie di*.

1. CORPORA UTILIZZATI. Dopo una cursoria disamina dei tre subcorpora specialistici NUNC di cucina, motori e fotografia, rivelatisi di dimensioni non sufficienti per l'analisi del pattern sintagmatico, si sono utilizzati corpora più consistenti, ossia:

(j) i NUNC (Newsgroups UseNet Corpora), serie multilingue di corpora basati su testi di UseNet e liberamente interrogabili online senza restrizioni (cfr. i contributi di Barbera ¶ 1, § 2.2.5, e Corino ¶ 13 in questo volume), ed in particolare i dati italiani di NUNC-IT Generic I+II, costituito da 276.795.236 token basati su una quindicina di mesi di post (grossomodo tra il settembre 2002 ed il gennaio 2004);

(ij) il CORIS (CORpus di Italiano Scritto) creato dal CILTA[4] e risultato di una ricerca svolta nell'àmbito dell'Università di Bologna; contiene 110 milioni di token ed è stato aggiornato nel luglio 2005 tramite un corpus di monitoraggio[5]. È costituito da una raccolta di testi autentici e ricorrenti nell'uso, selezionati come rappresentativi dell'italiano attuale[6].

2. IL PATTERN "TUTTO/A UN(A) __ DI". Come osservato nella premessa, questa ricerca intende prima di tutto verificare l'ipotesi che in italiano *esista* un pattern sintagmatico rappresentabile schematicamente come "tutto/a + un(a) + __ + di". A questo scopo si è proceduto lanciando le due query [1a] e [2a], eventualmente ricombinabili nella sola [3]:

[1a] `[word="tutta"][word="una"][pos="NOM"][word="di"]`
[2a] `[word="tutto"][word="un"][pos="NOM"][word="di"]`

[3] `[word="tutt[o|a]"][word="un|una"][pos="NOM"][word="di"]`

L'interrogazione per lemma `[lemma="tutto"]`, che sembrerebbe in prima istanza più efficace e metodologicamente più sensata, ha posto invece problemi a causa del consistente numero di forme plurali *tutti* / *tutte* in contesti che si sono poi rivelati difformi rispetto agli scopi della nostra ricerca, come [4b], anche per la frequenza di casi in cui l'aggettivo quantificativo appartiene ad un distinto costituente, come in [4c-d]. Si è pertanto preferita l'analisi sulle "*words*", verificando i contesti con nomi sia femminili sia maschili.

---

[4] Centro interfacoltà di linguistica teorica e applicata "L. Heilmann". Il corpus è accessibile previa registrazione a contesti limitati: gli esempi, infatti, ammettono al massimo 160 caratteri – neppure parole! – di contesto.

[5] Il corpus di monitoraggio, inglobato con cadenza biennale, è composto da 10 milioni di parole ripartite tra le varietà testuali nello stesso modo del CORIS. Copre gli anni dal 2001 al 2004 e l'etichetta che lo contraddistingue nello schema delle concordanze è "MON2001_04".

[6] I testi sono suddivisi in cinque macro-unità: *stampa* (38 milioni di parole), *narrativa* (25 milioni di parole), *prosa accademica* (12 milioni di parole), *prosa giuridico-amministrativa* (10 milioni di parole), *miscellanea* (10 milioni di parole), *ephemera* (testi a mano, a stampa e, principalmente, in formato elettronico, caratterizzati da una breve permanenza: 5 milioni di parole), a cui si sommano le 10 milioni di parole nel corpus di monitoraggio (cfr. nota 5). Da qui le sigle indicate negli esempi tratti dal CORIS (STAMPA, NARRAT, PRACC, PRGAMM, MISC, EPHEM), accompagnate di volta in volta da ulteriori sottospecificazioni.

[4a] `[lemma="tutto"][pos="DET.*"][pos="NOM"][word="di"]`

[4b] Erode , accortosi che i Magi si erano presi gioco di lui , s'infuriò e mandò ad uccidere **tutti i bambini di** Betlemme e del suo territorio dai due anni in giù , NUNC Generic I,

[4c] ragazzi quì bisogna farsi **tutti un esame di** coscienza ... NUNC Generic II,

[4d] Io consiglio a **tutti un portale di** ricerca che l'università di Torino ha inserito in rete . NUNC Generic I.

La prima query [1a], basata su *words* femminili, restituisce nei NUNC rispettivamente 809 e 750 occorrenze (distinguendo, come faremo sempre anche in seguito nel testo, i due sottocorpora NUNC Generic I e II), di cui circa l’80%, rispettivamente 679 e 588 occorrenze, è rappresentato da “tutta una serie di” [1b]. I restanti risultati mostrano invece un’elevata varietà lessicale [1c-h]:

[1b] Tu invece mi citi un sito che , trascurando il fatto che assomiglia in modo sinistro ad un sito Herbalife , sbrodola **tutta una serie di** strali , accuse e infamie contro i vaccini , dipingendo il mondo mondo con uno scenario da guerra nucleare : NUNC Generic II,

[1c] Il termine " Italia ", come si vede , è ricorrente (Servizio Italia , Italcantieri , e **tutta una sfilza di** Holding Italiana) , [...] . NUNC Generic I,

[1d] La cosa si può effettivamente notare , e dal treno di rotolamento , e dalle dimensioni proporzionalmente ridotte del carro . Tralasciando inoltre **tutta una miriade di** piccoli dettagli che non depongono per una assoluta fedeltà storica [...] . NUNC Generic I,

[1e] Meno accanito nel combattere il dolore , per cui assuefatto all ' idea della svolta epocale , che svolta senza ... mettere la freccia ed investe **tutta una pletora di** illusioni più o meno sbandierate . NUNC Generic I,

[1f] [...] le catastrofi che accadono nel mondo non sono mai la conseguenza o l' effetto di un unico motivo , d' una causa al singolare , ma sono come un vortice , un punto di depressione ciclonica nella coscienza del mondo , verso cui hanno cospirato **tutta una molteplicità di** causali convergenti Le motivazioni ci sono tutte . NUNC Generic I,

[1g] Certo che ad avere un testo come il ddj , con le varianti filologiche che ha , si dev ' essere per forza così . la sera della festa della divinità locale , c'era **tutta una schiera di** questi picchiatelli . NUNC Generic I,

[1h] Tenga conto pero' che esistono **tutta una gamma di** prodotti vita - che vendono anche le banche - che hanno costi ridottissimi , anche inferiori a quelli delle banche stesse . NUNC Generic I.

Poiché *serie* è risultato quantitativamente predominante nei dati estratti con la query indicata, abbiamo proceduto anche esplicitandone l’assenza, per velocizzare la ricerca delle “varianti” di *serie*:

[5] `[word="tutta"][word="una"][word!="serie"][word="di"]`

D’altra parte la maggiore frequenza di *serie* ci ha spinto a raffinare la ricerca su questo lessema incrociando i risultati delle due query seguenti:

[6] `[word="tutta"][word="una"][lemma="serie"]`

[7] `[word="tutta"][word="una"][lemma="serie"][word="di"]`

Abbiamo potuto osservare come i risultati della query [6] (Subset I: 698 occorrenze, 603 nel Subset II) siano non solo ovviamente restituiti anche dall'interrogazione completa [7], ma che i risultati di quest'ultima (Subset I: 679 occorrenze, Subset II: 588 occorrenze) si discostino di poco da quelli di [6]. La controprova [8a] con l'operatore di negazione *not* " ! ", inoltre, conduce esattamente a 19 e 15 occorrenze nei due rispettivi sottocorpora:

[8a] `[word="tutta"][word="una"][lemma="serie"][word!="di"]`

mostrando, oltre a più rare forme apostrofate (*tutta una serie d' informazioni riservate*, NUNC Generic I) e ad alcuni errori di battitura (*tutta una serie dieffetti nocivi*, NUNC Generic I-II), casi interessanti di modificazione, per lo più aggettivale [8b], del sostantivo *serie*, anche in forme cognitivamente "pesanti" come superlativi [8c], modificatori aggettivali a loro volta modificati come in [8d] ed incisi [8e]:

[8b] Però non vorrei che fosse l ' inizio di **tutta una serie** *infinita* di errori . NUNC Generic I,

[8c] In verita' non solo le assicurazioni stanno pesantemente incidendo sugli equilibri sociali degli italiani , bensi', c' e **tutta una serie** *intricatissima* di terziario e di intermediazione – che per brevita' non cito – che sta letteralmente soffocando la rigenerazione del reddito nel nostro Paese . NUNC Generic II,

[8d] Oltre ai pezzi tante volte eseguiti ed amati cmq dal pubblico , c' e' spazio poi per **tutta una serie** *altrettanto variabile* di brani del passato , [...] NUNC Generic II,

[8e] Al release dei pulsanti , una piccola animazione porta in scena **tutta una serie** *(per la precisione)* di altri pulsanti (tipo scritte-pulsanti) . NUNC Generic II.

Accanto a [8b]-[8e], altrettanto interessanti sono le modificazioni che precedono il sostantivo *serie*, per le quali si è operata un'ulteriore ricerca ([9a]) che ne isolasse in modo mirato la consistenza: Subset I: 57 occorrenze, Subset II: 34 occorrenze – di cui tuttavia un certo numero (16 e 15 rispettivamente) include casi diversi, disambiguabili solo manualmente[7], con altra struttura sintagmatica (cfr. [9b] in nota 7) o semantiche specifiche (cfr. [9c] in nota 7) di *serie*.

[9a] `[word="tutta"][]{2,3}[word="serie"]`

Sono dunque numericamente scarsi i casi interessanti di modificazione aggettivale, talvolta occorrenti anche con un dimostrativo, nei quali non sembra però sempre da escludersi la lettura di insieme indefinito che caratterizza *tutta una serie di*:

[10a] `[word="tutta"] [pos="DET.*" | pos="PRO.*" | pos="ADV"]? [pos="ADJ"] [lemma="serie"] [word="di"]`

[10b] L' attendibilita' di **tutta questa lunga serie** di " ha detto " , " ha affermato " ecc. del resto e' gia' di per se' rivelata dal riferimento alla camera Kirlian , la cui " aura " si e' visto che si manifesta [...] NUNC Generic I,

[10c] [...] : d' altra parte , non esiste alcuna evidenza , almeno per ora , che **tutta questa incredibile serie** di eventi abbia un' origine un po' meno naturale di quanto non si pensi. NUNC Generic II.

---

[7] Esempi significativi di questi risultati spuri possono essere:

[9b] Se in una campagna " avatarista " un personaggio / giocatore " vince " per **tutta** una serata una **serie** di sfide contro vari " mostri ", e poi alla fine muore , ha vinto o ha perso ? NUNC Generic II,

[9c'] NO !!! mi sto riguardando **tutta** la seconda **serie** , quella satanica , la crisi di Scully , tensione a mille , spettacolo . NUNC Generic I,

[9c''] Prendendo ovvero sia **tutta** gente di **serie** A coinvolgendoli nel progetto Florentia – (domani Fiorentina) . NUNC Generic II.

Per quanto non sia possibile verificare online la frequenza della sequenza non preceduta da *tutta*, quantomeno non in un'analisi quantitativa controllata[8], verifichiamo nondimeno l'eventuale consistenza di forme con determinanti diversi da *una*:

[11] `[word="tutta"][pos="DET.*"][lemma="serie"][word="di"]`

che restituisce nel Subset I: 768 occorrenze, nel subset II: 615 occorrenze.

La differenza di risultati rispetto a quelli restituiti da [7] non è rilevante (complessivamente 89 e 26 occorrenze di "tutta *la* serie di") ed indica piuttosto un alto grado di compattezza sintattica della costruzione, sia nella reggenza preposizionale, sia nella bassa frequenza delle pur possibili interpolazioni di ulteriore materiale linguistico (ess. [8b-e]), sia ancora nella predominanza del modificatore *tutto* e dell'articolo indeterminativo.

Nel complesso, la frequenza di *serie* amplifica notevolmente l'effetto generale di frequenza del pattern con *words* di genere femminile. La query con *words* maschili [2a] conduce infatti ad un numero decisamente inferiore, rispettivamente 101 e 100 occorrenze nei due subset:

[2b] I medici dovrebbero studiare di più in molti casi e non solo di psicologia , a volte volte è proprio la lor o materia ad essere carente ... avrei **tutto un repertorio di** racconti da fare . NUNC Generic I,

[2c] C'è **tutto un gruppo di** storie che riguardano la città : questioni di scala , distanza e spazi contestati . NUNC Generic I,

[2d] - era un misto di stupore barocco e di romanticismo , pathos , ,passione corale , espressione autentica dell' animo popolare , con la gioia dell' attesa , la spontaneità , la semplicità , la fantasia , il gusto della rappresentazione e il senso del mistero divino , **tutto un caleidoscopio di** sensazioni e emozioni che vengono rivissuti in un momento intensamente lirico malinconico e straziante , pur nella sua purissima dolcezza espressiva [.] NUNC Generic I,

[2e] In fondo , resto io il sognatore , quello con **tutto un campionario di** idee assurde che sembrano sempre non funzionare , nonostante la creatività che cerco di infondere loro. NUNC Generic I,

[2f] Te lo sconsiglio perche' l' oggetto Array ogni volta che ne viene creato uno , si porta dietro **tutto un malloppo di proprieta'** e metodi che sono del tutto superflui nell dei casi come il tuo . NUNC Generic I,

[2g] Perchè per ricevere una mail devo beccarmi **tutto un minestrone di** vaccate colorate nelle quali quello che c' è da leggere è quasi sepolto ? NUNC Generic I.

Il quadro che emerge dai risultati dei NUNC trova conferme nell'analisi parallela del CORIS, in cui si ha per altro una distribuzione quantitativa ancor più consistente, in relazione alla più ridotta ampiezza del corpus: 677 occorrenze complessive del pattern *tutta una serie di* su un totale di 755 risultati per la query con nome femminile, con una presenza trasversale nelle diverse varietà testuali che ci mostra una combinazione sintagmatica comune in tutte le aree considerate, dalla stampa alla narrativa, fino alla prosa accademica e giuridico-amministrativa[9]:

---

[8] Quantitativamente, infatti, i risultati della query `[word="una"][lemma="serie"][word="di"]` ed a maggior ragione di `[pos="DET.*"][lemma="serie"][word="di"]` eccedono in entrambi i subset la soglia dei 1000 risultati per ora tecnicamente visualizzabili dall'interfaccia web dei NUNC.

[9] Anche per il CORIS vale una restrizione tecnica: il limite di soli 300 risultati supportati dall'interfaccia. Viene, ossia, fornito il numero esatto di occorrenze presenti nel corpus, ma è possibile visualizzarne solo 300, che abbiamo giocoforza considerato nella trattazione che segue come campione rappresentativo del fenomeno.

[12a] [Arrivato ai 65] in forma smagliante , potrò dedicarmi completamente al riposo ed a **tutta una serie di** piccoli ma fantastici passatempi trascurati da sempre .[10] CORIS - NARRATRacc,

[12b] [Per quanto riguarda la nostr]a realtà , mi aspetto una soluzione positiva da parte dell ' Asi di **tutta una serie di** problematiche sui lotti industriali e sulle aree di competenza del consor[zio .] CORIS - STAMPAQuot,

[12c] [Alcuni preparati di mitocondri , isolati per centrifugazione , si sono rivelati ab]erranti e inutilizzabili per lo studio biochimico perché contenenti **tutta una serie di** enzimi idrolizzanti [...] [.] CORIS - PRACCVolum,

[12d] Dal 2 gennaio 2000 le Preture non esistono più . **Tutta una serie di** reati di competenza pretorile col nuovo anno sono passati ai tribunali . CORIS - PRGAMMRivi.

Anche per quanto riguarda la proporzione tra *serie* e gli altri lessemi che possono occupare il segnaposto, la distribuzione in testi di italiano scritto è dunque quantitativamente simile a quella emersa nei newsgroups: sui primi 300 casi analizzabili (cfr. nota 9) nel CORIS solo il 20% degli esempi non contiene il sostantivo *serie*, mostrando comunque un certo tasso di libertà nella selezione lessicale del segnaposto, sia nella query basata su lessemi femminili,

[13a] Penso **a tutta una sequenza di** parole da dire , che si complicano al punto che mi esce solo [: - Ti fa male ?] CORIS - NARRATRoma,

[13b] [Ha vinto il savoiardo Joel Chenal , che vive poco oltre il Piccolo San Bernardo , a La Rosiere , uno che in vita o che in vita sua] non era mai andato più in là di un quinto posto : battendo Hermann Maier , e **tutta una sfilza di** campioni illustri , da Aamodt a Von Gruenigen (che delusione , decimo) . CORIS - STAMPAQuot,

sia nella query con nomi maschili (246 risultati):

[14a] Altro aspetto , secondo me indicativo è rappresentato da **tutto un insieme di** costosi corredi canini e di un certo tipo di alimentazione [con carne o pesce di qualità che viene acquistata solo per la nutrizione dell ' animale e questo mi pare inaccettabile .] CORIS - NARRATVari,

[14b] Ma sullo sfondo si intuisce **tutto un complesso di** allegorie e di significati riposti [che ripetute volte si è tentato di penetrare , probabilmente invano .] CORIS - PRACCVolum.

Il complesso di questi dati conferma dunque la preponderanza quantitativa del lessema *serie* sulle altre varianti del pattern sia nei NUNC sia nei testi di italiano scritto più sorvegliato del CORIS. D'altra parte, nonostante la prevalenza di *serie*, esiste in entrambi i corpora un ampio ventaglio di altre possibilità di riempimento del pattern sintagmatico, come si evince dalle Tav. 1a ed 1b, in cui condensiamo i risultati dei venti segnaposto più frequenti nel pattern "tutto/a + un(a) + __ + di" nei due corpora, riportando il numero dei token che vi rientrano come segnaposto più frequenti[11]:

[10] Gli esempi tratti dal CORIS sono riportati con tutto quello che di contesto pertinente è preso dalla finestra di 160 caratteri, limite massimo di ricerca concesso per il corpus. Nei molti casi in cui il limite rendeva l'esempio del tutto infruibile, abbiamo dovuto integrare il contesto ricorrendo ad un collage con ulteriori query *ad hoc*, il cui risultato è riportato negli esempi tra parentesi quadre.

[11] Il diverso (meno parametrizzabile) sistema di interrogazione del CORIS, tramite la query "tutto+un+*+di", risalirebbe a numerosi casi che vedono * = *verbo*,

NUNC		CORIS		NUNC		CORIS	
serie	1267[12]	serie	240[13]	mondo	10	insieme	13
gamma	14	schiera	5	sistema	5	sistema	10
sfilza	7	gamma	4	complesso	4	repertorio	6
miriade	5	rete	3	universo	4	complesso	4
rete	5	fila	2	campionario	3	mondo	4
massa	4	fioritura	2	coacervo	2	concerto	3
categoria	3	gerarchia	2	coro	2	corteo	3
fetta	3	categoria	1	fascio	2	filone	3
somma	3	congerie	1	gruppo	2	gruppo	3
classe	2	covata	1	miscuglio	2	ventaglio	3
complessità	2	lista	1	repertorio	2	viavai	3
distesa	2	nidiata	1	caleidoscopio	1	assortimento	2
fila	2	matassa	1	casino	1	campionario	2
flora	2	miriade	1	corollario	1	seguito	2
miniera	2	moltitudine	1	cozzo	1	andirivieni	1
pappardella	2	pluralità	1	florilegio	1	armamentario	1
ridda	2	polifonia	1	guazzabuglio	1	bagaglio	1
schiera	2	sfilza	1	malloppo	1	carrozzone	1
sequela	2	turba	1	mare	1	esercito	1
tipologia	2	varietà	1	mucchio	1	universo	1

Tav. 1a. "tutta una X di". Tav. 1b: "tutto un X di".

Interessante notare come i segnaposto restituiti dai corpora siano coperti per il 100% da nomi indicanti quantità, o meglio insiemi di oggetti (tutto un *gruppo*, un *complesso*, un *campionario*, un *caleidoscopio*, tutta una *miriade*, una *massa*, una *gamma*, ecc.); intervengono solo sfumature semantiche di distribuzione interna dell'insieme, più o meno omogeneo (cfr. *campionario* vs *massa*). La varietà di segnaposto possibili confortata dai dati a nostra disposizione sembra inoltre indicare una cristallizzazione non solo di *tutta una serie di*, quanto piuttosto dello scheletro che ne è costitutivo - per quanto si possano comunque ipotizzare forme originariamente più frequenti da cui il pattern stesso è nato[14].

---

[15] [Allora , Lippi , dal quel fatidico 5 dicembre ' 95 , giorno in cui lei passava per i corridoi di Milano 2 e , senza neppure provare , fu bu]ttato in studio per sostituire Teocoli a " Mai dire gol " , la sua carriera è **tutto un <u>fiorire</u> di** successi , proposte , richieste .

CORIS - STAMPAQuot

esclusi dal conteggio nei NUNC attraverso la specificazione `[pos="NOM"]` e non presi in considerazione in questa fase della ricerca, malgrado si possa sin d'ora registrare per la maggior parte delle forme verbali una semantica compatibile con la quantificazione su insiemi che caratterizza anche i nomi presenti nel pattern (tutto un *fiorire, proliferare, brulicare, pullulare* di).

[12] Calcolati su entrambi i subset NUNC Generic I (679 occorrenze) e II (588).

[13] Su 300 risultati totali (cfr. nota 9).

[14] La verifica dell'eventuale presenza di *tutta una serie di* nella stampa periodica milanese dell'Ottocento (Bonomi - De Stefanis Ciccone - Masini 1983) ha voluto muoversi proprio in questa direzione. I dati non ne dimostrano però ancora l'esistenza; tra le concordanze possiamo solo menzionare casi del tipo (i contesti sono fissi ad una riga, senza contesto pieno, che in almeno un caso abbiamo pensato meglio di integrare):

[16a] [Faremo] gustare a' nostri lettori **una serie di** queste lettere dell'Agatocle

La compattezza semantica dei dati presentati (*supra*, Tav. 1) dimostra la selettività del pattern suggerendo la compresenza di tratti morfosintattici (corrispondenti alla sequenza sintagmatica stessa ed alle sue caratteristiche strutturali) e semantici (la denotazione di insiemi quantificati) che sarebbe difficile considerare casuale.

Se si tratta di un pattern con caratteristiche strutturalmente prevedibili e ricorrenti, è dunque lecito chiedersi se una struttura di questo tipo non debba avere una qualche rilevanza grammaticografica che permetta di ascriverlo ad una specifica parte del discorso non corrispondente a nessuna delle parti del discorso dei suoi componenti (*tutto/a* = Aggettivo; *un/una* = Determinante; *X* = Nome; *di* = Preposizione), e neppure determinabile composizionalmente dalla somma dei singoli costituenti. In effetti lo statuto categoriale di un pattern come "tutto/a un(a) X di" è un punto particolarmente controverso della descrizione grammaticale di diverse lingue.

Nel § 3 passeremo in rassegna alcune delle proposte correnti per strutture sintagmatiche simili a quella qui considerata. Nel § 4 mostreremo poi come la questione dello statuto categoriale del pattern possa essere in parte chiarito approfondendo la ricerca in una prospettiva *corpus-based* che studi le caratteristiche strutturali del pattern considerando anche fattori di differenziazione diafasica o diamesica tra corpora diversi[15].

3. LO STATUTO CATEGORIALE DI "TUTTO/A UN(A) __ DI". La complessa natura delle strutture esaminate è dimostrata dal forte grado di discrepanza interpretativa che si può riscontrare nella letteratura sull'argomento, soprattutto se si prendono in considerazione tradizioni grammaticografiche di lingue diverse. Per quanto diversificate, le interpretazioni proposte possono comunque essere ricondotte a due filoni principali, che vedono opporsi definizioni prevalentemente od esclusivamente semantiche ad altre di natura piuttosto morfosintattica.

Il carattere semantico di alcune definizioni è intrinseco alle stesse scelte terminologiche che puntano sul significato quantificazionale di nomi come *serie*, *insieme*, *complesso*, *gruppo* etichettati come «noms de quantité (indéterminée)» (Flaux 2001, p. 155) o come «collectives» (Michaux 1992). A questo proposito Bosque 1999, pp. 23-26, fa notare come il termine "collettivo" dovrebbe riferirsi in senso stretto solo a lessemi di forma singolare ma con referenza plurale (sp. *arboleda*, *vecindario*, it. *esercito*, *mobilia*), caratterizzandosi quindi come una specificazione semantica della categoria grammaticale del numero (cfr. Gil 1996, pp. 66-70, e Corbett 2000). D'altra parte la soluzione terminologica adottata da Bosque 1999, p. 18, che parla di «sustantivos cantificativos», preferendo questa etichetta a «nombres de medida», è però ugualmente insoddisfacente in quanto considererebbe il pattern "tutto/a un(a) __ di" come istanza della categoria morfosintattica dei nomi, tralasciando il fatto, segnalato peraltro da Bosque 1999, p. 26, che si tratta invece di determinanti di nomi, cioè di modificatori e non di nomi veri e propri.

Su questa natura morfosintattica di determinanti insiste infatti la linguistica francese di ispirazione lessico-grammaticale proponendo quindi etichette come «déterminants nominaux»

[16b] con **tutta la lunga serie di** accidentali disastri che giornalmente

[16c] di là **quella serie infinita di** ridicole pretensioni e di esagerate

[15] Al fine di verificare la rispondenza dei risultati dei NUNC e CORIS a varietà diamesiche non scritte, avevamo anche ipotizzato un confronto con corpora di lingua parlata, servendoci in particolare di due risorse:

(j) BADIP (Banca Dati dell'italiano parlato), sito gratuito dedicato alla pubblicazione di corpora e altri materiali per l'analisi e lo studio della lingua italiana parlata; contiene una versione online del LIP, il *Lessico di frequenza dell'italiano parlato* (De Mauro et alii 1993);

(ij) LABLITA, il Laboratorio Linguistico del Dipartimento di Italianistica dell'Università di Firenze, che si occupa della raccolta e gestione di corpora di parlato spontaneo (Cresti 2000). [Per l'accesso a questi dati ringraziamo Alessandro Panunzi].

Questi corpora di italiano parlato hanno tuttavia mostrato percentuali minime e poco rappresentative nella presenza di *tutta una serie di*, con un numero di risultati troppo circoscritto per un'analisi comparabile con quella operata su CORIS e NUNC (12 risultati nel LIP e 10 restituiti da LABLITA).

(Dessaux 1976), che colgono il carattere morfosintatticamente duplice di queste strutture, in parte nomi, in parte determinanti. Ad es., la sequenza sintagmatica *un paquet de voitures* viene definita come espressione della categoria dei «déterminants complexes composés figés» (Buvet 2001), che si oppongono primariamente ai «déterminants complexes composés non figés» (*beacoup de voitures*) e secondariamente ai «déterminants simples» (*cette voiture*, *une voiture*).

Una posizione di mediazione tra la prospettiva morfosintattica che insiste sulla natura di determinanti del nome ed una prospettiva semantica, che punta invece sulla natura quantificazionale, è rappresentata dall'ipotesi di considerare *un paquet de*, *una serie di*, *una manada de* all'interno della categoria dei quantificatori[16]. In questa prospettiva si pone un suggerimento di Petőfi 1979, sviluppato da Eikmeyer 1980, p. 97, e da Marello 1980 per l'italiano, che parlano di *Quantorspezifikatoren* e più specificamente di *idiomatische Quantoren* per sequenze come *un mucchio*, *un sacco di problemi* (Marello 1980, pp. 58-60)[17]. Infatti già Lyons 1977, § 11.4, richiamava l'associazione tra determinanti e quantificatori, assumendo implicitamente che la categoria dei quantificatori sia definibile come un'interfaccia tra la morfosintassi di un determinante e la semantica di termine quantificazionale. In ciò che segue sfrutteremo i corpora a nostra disposizione per verificare se ci offrono argomenti rispetto allo statuto categoriale del pattern individuato; in particolare cercheremo di verificare se sia perseguibile l'ipotesi interpretativa che li considera dei quantificatori proponendo quindi una possibile mediazione tra un'interpretazione semantica ed il riconoscimento del loro ruolo morfosintattico di determinanti.

Prima di passare all'analisi dei dati dobbiamo però ricordare un'altra proposta interpretativa che considera sequenze sintagmatiche del tipo di *a bunch of*, *a box of*, *a number of*, *a group of* come dei classificatori (Lehrer 1986). Questa ipotesi si basa sul fatto che un pattern come "tutto/a un(a) __ di" mostra in effetti restrizioni di selezione rispetto ai nomi di cui si esplicita l'appartenenza ad un insieme: la più ovvia restrizione implica che ad esempio *una serie di* possa quantificare solo su nomi plurali numerabili, e non su singolari collettivi (*una serie di persone* vs **una serie di gente*) a differenza di *un sacco di* (*un sacco di persone* / *un sacco di gente*). Restrizioni sulla numerabilità del quantificato riguardano però anche veri e propri quantificatori (*much* e *many* in inglese, cfr. Gil 2001, p. 1275) e non sono quindi proprie solo dei classificatori[18].

Si deve inoltre tener presente che un classificatore prototipico è esemplificato dal sintagma nominale ungherese in [17] che dimostra come il classificatore numerale *szál* si accompagni al quantificatore (*egy*) senza sostituirlo, mentre sequenze sintagmatiche come *una serie di*, *un mucchio di*, ecc. non richiedono (e non permettono) altri elementi di quantificazione essendo già di per sé dei quantificatori che denotano insiemi di oggetti:

[17]	*egy*	*szál*	*gyertya*	'one candle'	
	one	CL:LONG:CYLINDRIC	candle		Aikhenvald 2000, p. 102.

Concordiamo dunque con l'invito di Aikhenvald 2000 a non estendere troppo la nozione di classificatore, ma rileviamo anche che l'idea di considerare strutture del tipo di "tutto/a un(a) __

[16] Questa soluzione è del resto suggerita anche da Bosque 1999, p. 24, e non a caso le strutture qui analizzate, oltre ad essere trattate nel capitolo dedicato al «nombre común» vengono poi cursoriamente citate anche nel capitolo sui quantificatori (Sánchez López 1999, p. 1050). Più esplicito il riconoscimento della particolare natura di queste strutture come nomi quantificativi che assumono funzione di quantificatori nel trattamento proposto per il catalano da Martí Girbau 2002, pp. 1301-1302, e soprattutto da Brucart - Rigau 2002, pp. 1542-1543.

[17] In linea di principio i *Quantorspezifikatoren* sono sempre combinabili con un quantificatore. Questo permette di distinguere tra *Quantorspezifikator* ed *idiomatischer Quantor*: *un mucchio di* può essere impiegato come quantificatore idiomatico nell'accezione in cui non co-ricorre con quantificatori (**due/molti mucchi di problemi* ["idiomatischer Quantor"] vs. *due mucchi di spazzatura* ["Quantorspezifikator"], cfr. Marello 1980, pp. 58-60).

[18] Cfr. anche i «numeral classifiers» in Gil 2005, § 1 e 4, intesi come «sortal numeral classifiers» che dividono i nomi numerabili in classi semantiche, escludendo i «mensural numeral classifiers» del tipo *one glass of water, two pounds of sand*, presenti in quasi tutte le lingue.

di" come classificatori rappresenta un tentativo di superare la dicotomia tra interpretazione semantica ("nomi di quantità", "collettivo", ecc.) e morfosintattica ("determinanti"), attribuendoli ad una categoria grammaticale, che, non diversamente dai quantificatori, sia interpretabile come l'interfaccia tra morfosintassi e semantica. Anche in questo caso si può ricordare il richiamo di Lyons 1977, § 11.4, alla relazione categoriale non solo tra determinanti e quantificatori, ma anche tra determinanti, quantificatori e classificatori, con particolare riguardo alla specifica relazione tra quantificatori e "mensural classifiers" da un lato ed a quella tra determinanti e "sortal classifiers" dall'altro.

È quindi ipotizzabile che lo statuto categoriale delle sequenze sintagmatiche qui analizzate possa anche permettere un'analisi diversificata tipologicamente a seconda della lingua che si sta descrivendo, per cui in alcune lingue le stesse strutture potrebbero condividere più tratti con i classificatori che con i quantificatori.

Nel paragrafo seguente comunque cercheremo conferme empiriche all'ipotesi che abbiamo considerato teoricamente più soddisfacente, e cioè che "tutto/a un(a) __ di" in italiano sia interpretabile come un quantificatore.

4. CONFRONTI TRA CORPORA. Per verificare le ipotesi sullo statuto categoriale del pattern sintagmatico "tutto/a un(a) __ di" abbiamo scelto di approfondire le caratteristiche morfosintattiche di *tutta una serie di*, che, come ampiamente dimostrato dai risultati quantitativi (cfr. § 2), è il più frequente tra le possibili repliche del pattern individuato.

Date le opzioni interpretative presentate nel § 3, abbiamo considerato il tratto morfosintattico dell'accordo di numero tra sintagma nominale soggetto e verbo. Osservando i dati di corpora si evince infatti che "tutta una serie di" in funzione di soggetto ammette sia l'accordo al singolare [18a] che al plurale [18b],

[18a] [[...] il tutto incastonato con decorazioni minuziose di vetri , lapi]slazzuli , cristallo , oro e argento dai riflessi abbaglianti . C ' *era* anche **tutta una serie di** oggetti ornamentali . CORIS – NARRATTrRo,

[18b] Star Trek ha dato vita al fenomeno dello slash , dopo tutto . Ma c ' *erano* **tutta una serie di** indizi che portavano in quella direzione . NUNC Generic I,

anche con equivalente contenuto proposizionale espresso nei due modi possibili:

[19a] Berlusconi ha garantito che andrà " fino in fondo " nella vicenda affermando come nel processo si *sia verificata* **tutta una serie di** situazioni come , per esempio , " la mancata escussione di testi importantissimi " NUNC Generic II,

[19b] Nel processo , ha proseguito Berlusconi , « si *sono verificate* **tutta una serie di** situazioni come la mancata escussione di testi importantissimi » . NUNC Generic II.

Considerato che il nome *serie* è di per sé un singolare, la presenza di accordi *ad sensum* al plurale è stato riconosciuto come un tratto caratterizzante di queste strutture "pseudopartitive" (Martí Girbau 2002, p. 1288; Brucart - Rigau 2002, p. 1535), ma più in generale è anche interpretabile come il sintomo di un processo di decategorializzazione del nome, che perde i suoi tratti morfosintattici di singolare. Ora, riprendendo la discussione presentata nel § 3, possiamo osservare che la perdita dei tratti categoriali di nome sembra difficilmente compatibile con un'interpretazione grammaticografica come "nome collettivo", "nome di quantità", "termine mensurale", ma anche "nome quantificazionale" *à la* Bosque 1999, che classificherebbe *serie* in ogni caso come appartenente alla categoria "nome". La decategorializzazione fa propendere piuttosto per un'interpretazione come modificatore del nome che mantiene però una semantica

quantificazionale nel riferimento ad insiemi indefiniti e che è quindi interpretabile come una forma di quantificatore. La decategorializzazione è uno dei fenomeni che insieme ad altri partecipano ai processi di grammaticalizzazione (Heine et al. 1991; Lehmann 2002) e ciò permetterebbe di interpretare il pattern “tutto/a un(a) __ di”, almeno quando si realizza nella forma *tutta una serie di* come un quantificatore in corso di grammaticalizzazione. Il trasferimento del tratto [+plurale] dal sostantivo al modificatore *serie* è già da solo indizio di un processo in corso, non necessariamente divergente da quel procedimento metonimico che, per Diewald 1997, sposterebbe il focus cognitivo del parlante dal gruppo ai componenti discreti che ne fanno parte.

L’ipotesi di un processo di decategorializzazione ancora in corso, e di conseguenza sottoposto a forte variabilità in dipendenza da parametri sociolinguistici, è anche confermata da un confronto tra i due corpora, che mostra un più frequente accordo “scorretto” al plurale nei testi di newsgroup e percentuali più alte di accordo al singolare negli scritti contenuti nel CORIS, formalmente più controllati. Il quadro riassuntivo si presenta come da Tav. 2:

	Totale	*Accordo SG*	*Accordo PL*	*% verbi SG*	*% verbi PL*
NUNC	156	64	92	41,02	58,79
CORIS	46	27	19	58,69	41,30

Tav. 2: Accordo di numero soggetto/verbo.

Il chiasmo di cifre risultante nelle percentuali è indicativo di una tendenza diversificata nei tipi di testo che i due corpora rappresentano: i newsgroup, cronologicamente più recenti e generalmente frutto di una produzione linguistica più “spontanea” (cfr. Corino ¶ 13 *infra*), potrebbero confermare una tendenza emergente, non ancora consolidatasi nei generi più tradizionali di italiano scritto.

5. CONCLUSIONI. In questo lavoro abbiamo dimostrato come la sequenza “tutto/a un(a) __ di” rappresenti un pattern sintagmatico che copre una classe naturale di strutture, di cui nella seconda parte abbiamo verificato la rilevanza grammaticografica ipotizzando che si tratti di un quantificatore in corso di grammaticalizzazione.

Questo punto è stato dimostrato considerando la diversa distribuzione dei fenomeni di accordo tra soggetto e verbo in corpora di diversa natura: la preponderanza di contesti con forme plurali è chiaro indicatore di una decategorializzazione *in itinere* del pattern preso in esame, il cui impatto sul sistema complessivo dei quantificatori ci proponiamo di indagare in future ricerche, rivolte in particolare ad approfondire la funzione semantica di *tutta una serie di* rispetto ad altri mezzi di quantificazione.

BIBLIOGRAFIA.

AIJMER - ALTENBERG

1991 *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, edited by Aijmer Karin and Altenberg Bengt, London, Longman, 1991.

AIKHENVALD

2000 Alexandra Aikhenvald, *Classifiers. A Typology of Noun Categorization Devices*, Oxford, Oxford University Press, 2000 “Oxford Studies in Typology and Linguistic Theory”.

BARBERA

¶ 1. Manuel Barbera, *Tra bmanuel.org e corpora.unito.it. Per la storia di un gruppo di ricerca*, in questo volume, pp. 3-20.

BLANCO et alii
2001 Xavier Blanco - Pierre-André Buvet - Zoé Gavriilidou, *Détermination et formalisation*, Amsterdam - Philadelphia, John Benjamins, 2001.

BONOMI - DE STEFANIS CICCONE - MASINI
1983 *La stampa periodica milanese della prima metà dell'Ottocento: testi e concordanze*, a cura di Ilaria Bonomi, Stefania De Stefanis Ciccone e Andrea Masini, Pisa, Giardini, 1983.

BOSQUE
1999 Ignacio Bosque, *El nombre común*, in BOSQUE - DEMONTE 1999, vol. 1, pp. 3-75.

BOSQUE - DEMONTE
1999 *Gramática descriptiva de la lengua española*, dirigida por Ignacio Bosque y Violeta Demonte, preámbulo de Fernando Lázaro Carreter, índices a cargo de Ma. Victoria Pavón Lucero, Madrid, Espasa-Calpe, 1999, 3 voll.

BRUCART - RIGAU
2002 Josep M. Brucart - Gemma Rigau, *La quantificació*, in SOLÀ et alii 2002, pp. 1517-1589.

BUVET
2001 Pierre-André Buvet, *Les déterminants intensifs*, in BLANCO et alii, 2001, pp. 101-113.

CORBETT
2000 Greville Corbett, *Number*, Cambridge, Cambridge University Press, 2000 "Cambridge Textbooks in Linguistics".

CORINO
¶ 13. Elisa Corino, *NUNC (Newsgroup UseNet Corpora). Aspetti della testualità e questioni metodologiche*, in questo volume, pp. 225-252.

CORINO - MARELLO - ONESTI
2006 *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 | Proceedings of the XII EURALEX International Congress. Torino, Italia, 6th-9th September 2006*, a cura di Elisa Corino, Carla Marello e Cristina Onesti, 2 voll., Alessandria, Edizioni dell'Orso, 2006.

CRESTI
2000 Emanuela Cresti, *Corpus di italiano parlato*, 2 voll., Firenze, Accademia della Crusca, 2000.

DE MAURO et alii
1993 Tullio De Mauro - Federico Mancini - Massimo Vedovelli - Miriam Voghera, *Lessico di frequenza dell'italiano parlato*, Milano, ETASLIBRI, 1993.

DESSAUX
1976 Anne-Marie Dessaux, *Déterminants nominaux et paraphrases prépositionnelles: problèmes de description syntaxique et sémantique du lexique*, in "Langue Française" XXX (1976) 44-62.

DIEWALD
1997 Gabriele Diewald, *Grammatikalisierung: eine Einführung in Sein und Werden grammatischer Formen*, Tübingen, Niemeyer, 1997 "Germanistische Arbeitshefte 36".

EIKMEYER
1980 Hans-Jürgen Eikmeyer, *Quantoren, Quantormodifikatoren und Quantorspezifikatoren: Aspekte ihrer Explikation*, in EIKMEYER - JANSEN 1980, pp. 97-122.

EIKMEYER - JANSEN

1980 *Objektargumente, Grundelemente der semantischen Struktur von Texten*, herausgegeben von Hans-Jürgen Eikmeyer und Louise M. Jansen, Hamburg, Helmut Buske Verlag, 1980 "Papiere zur Textlinguistik / Papers in Textlinguistics", vol. III.

FLAUX

2001 Nelly Flaux, *Le classement des noms des quantité*, in BLANCO et alii 2001, pp. 151-161.

GIL

1996 David Gil, *Maltese 'Collective Nouns': A Typological Perspective*, in "Rivista di Linguistica" VIII (1996) 53-87.

2005 David Gil, *Numeral Classifiers*, in HASPELMATH et alii 2005, pp. 226-229.

GUENTHNER - SCHMIDT

1978 *Formal Semantics and Pragmatics for Natural Languages*, edited by Franz Guenthner and Siegfried J. Schmidt, Dordrecht, Reidel, 1978. [Essays in this collection are the outgrowth of a workshop held in June 1976].

HASPELMATH et alii

2005 *World Atlas of Language Structures*, edited by Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie, Oxford, Oxford University Press, 2005.

HEINE - CLAUDI - HÜNNEMEYER

1991 Bernd Heine - Ulrike Claudi - Friederike Hünnemeyer, *Grammaticalization: A conceptual framework*, Chicago,University of Chicago Press, 1991.

HUNSTON - FRANCIS

2000 Susan Hunston - Gill Francis, *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*, Amsterdam - Philadelphia, John Benjamins, 2000.

LEHMANN

2002 Christian Lehmann, *Thoughts on Grammaticalization*. Second, revised edition, Erfurt, Seminar für Sprachwissenschaft der Universität, 2005 "ASSidUE" 9, online come `http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-2058/ASSidUE09.pdf`

LEHRER

1986 Adrienne Lehrer, *English Classifier Constructions,* in "Lingua" LXVIII (1986) 109-148.

LEONI - DI BLASI

1981 *Lessico e semantica. Atti del XII Congresso Internazionale di studi della Società di Linguistica Italiana (Sorrento, 19-21 maggio 1978)*, a cura di Federico Albano Leoni e Nicola Di Blasi, Bulzoni, Roma, 1981.

LYONS

1977 John Lyons, *Semantics*, 2 voll., Cambridge - London - New York - Melbourne, Cambridge University Press, 1977.

MARTÍ GIRBAU

2002 Núria Martí i Girbau, *El SN: els noms*, in SOLÀ et alii 2002, pp. 1281-1335.

MARELLO

1980 Carla Marello, *Sprachspezifische Quantorspezifikatoren*, in EIKMEYER - JANSEN 1980, pp. 43-62.

1981 Carla Marello, *Per un pugno di parole: specificatori di quantificatori,* in LEONI - DI BLASI 1981, pp. 293-310.

MICHAUX
1992 Christine Michaux, *The collectives in French: A linguistic investigation*, in "Linguisticae Investigationes" XVI (1992) 99-124.

PETŐFI
1978 Petőfi János S[ándor], *Structure and Function of the Grammatical Component of the Text-structure World-structure Theory*, in GUENTHNER - SCHMIDT 1978, pp. 303-338.

RENOUF - SINCLAIR
1991 Antoinette Renouf - John Sinclair, *Collocational Frameworks*, in AIJMER - ALTENBERG 1991, pp. 28-43.

ROSSINI FAVRETTI
2000 *Linguistica e informatica.* Corpora*, Multimedialità e percorsi di apprendimento*, a cura di Rema Rossini Favretti, Roma, Bulzoni Editore, 2000.
2000a *Progettazione e costruzione di un* corpus *di italiano scritto: CORIS/CODIS*, in ROSSINI FAVRETTI 2000, pp. 39-56.

SÁNCHEZ LÓPEZ
1999 Cristina Sánchez López, *Los cuantificadores: Clases de cuantificadores y estructuras cuantificativas*, in BOSQUE - DEMONTE 1999, vol. 1, pp. 1025-1128.

SOLÀ et alii
2002 Joan Solà - Maria-Rosa Lloret - Joan Mascaró - Manuel Pérez Saldanya, *Gramàtica del català contemporani,* Barcelona, Empúries, 2002.

CORPORA E SITI DI RIFERIMENTO[19].

BADIP	http://languageserver.uni-graz.at/badip/badip/home.php
CILTA	http://www.cilta.unibo.it/
CORIS	http://corpora.dslo.unibo.it/coris_ita.html
LABLITA	http://lablita.dit.unifi.it
LIP	http://languageserver.uni-graz.at/badip/badip/20_corpusLip.php
NUNC	http://www.bmanuel.org/projects/ng-HOME.html

[19] I link sono stati aggiornati al 18/02/2007.

LUCA VALLE
(Università di Torino)

16. Ricerche su anglismi nei NUNC francesi ed italiani. *Tra "lurker", "lurkeur" ed altri prestiti.*

0. INTRODUZIONE**.** In questo lavoro presenterò alcuni risultati *in itinere* di una ricerca dedicata agli anglismi in francese ed in italiano[1], effettuata con l'ausilio dei corpora NUNC[2] elaborati all'interno del gruppo di ricerca di cui faccio parte[3].

In prima battuta, i corpora NUNC da me utilizzati per questa ricerca, sono stati i corpora specialistici NUNC-IT Cucina e NUNC-IT Motori per quanto riguarda la lingua italiana, e, in seconda battuta, i rispettivi corpora specialistici francesi del medesimo settore NUNC-FR Cucina e NUNC-FR Motori[4].

Nel prossimo paragrafo, darò brevemente alcune informazioni sui corpora che ho utilizzato finora. Nel paragrafo 2, spiegherò come ho effettuato l'estrazione degli anglismi dai corpora in oggetto. Nel paragrafo 3, presenterò i risultati ottenuti e nel paragrafo 4, tratterò di alcuni anglismi reperiti nei corpora italiani e francesi.

1. I CORPORA NUNC UTILIZZATI PER QUESTA RICERCA. Come già accennato nel paragrafo precedente, i corpora che ho utilizzato per la mia ricerca, sono per il momento quattro, e sono liberamente interrogabili al sito www.corpora.unito.it.

Il corpus NUNC-IT Cucina è un corpus specialistico di testi tratti da newsgroup italiani relativi al settore dell'alimentazione. Al suo interno sono compresi anche testi su argomenti correlati come ad esempio la ristorazione. Il numero di token è pari a 4.161.627, quello di type a 187.544 ed il numero di lemmi a 23.543.

Il corpus NUNC-IT Motori è un corpus specialistico di testi tratti da newsgroup italiani di motori, compresi anche testi su argomenti correlati, come ad esempio il mercato automobilistico. Il numero di token è 7.909.608, quello di type 273.744 ed il numero di lemmi 23.964.

I rispettivi corpora specialistici francesi, NUNC-FR Cucina e NUNC-FR Motori, presentano le stesse caratteristiche dei corpora italiani. Per quanto riguarda il corpus NUNC-FR Cucina, il numero di token è pari a 4.900.590, quello di type a 135.746 e quello di lemmi a 23.821. Per quanto riguarda il corpus NUNC-FR Motori, il numero di token è pari a 8.684.354, quello di type a 194.377 e quello di lemmi a 24.846.

---

[1] Avviata all'interno del Dottorato di ricerca in Linguistica, Linguistica Applicata, Ingegneria Linguistica (Ciclo XIX) con il titolo *Allestimento di corpora di newsgroup italiani e francesi. Estrazione, analisi e confronto di anglismi in francese e in italiano*, tuttora in corso.

[2] NUNC (Newsgroups UseNet Corpora) è una raccolta multilingue di corpora generici e specialistici composti da testi provenienti da newsgroup. I newsgroup sono forum telematici a libero accesso, liberamente disponibili su Internet, in cui ogni utente può partecipare alla discussione sull'argomento prefissato inviando un messaggio. Per maggiori dettagli rimando all'articolo di Barbera ¶ 1, § 2.2.5, e Corino ¶ 13, in questo volume. Basti qui ricordare che la scelta di creare corpora basati su testi provenienti da newsgroup ha comportato innumerevoli vantaggi, tra cui quello di fornire grandi quantità di dati per indagini dal punto di vista terminologico e lessicografico. Questo aspetto ha controbilanciato alcuni svantaggi (ad esempio l'abbondanza di testo ripetuto dovuto alla pratica del quoting) i quali sono stati in parte superati adottando particolari strategie di trattamento dei testi.

[3] FIRB 2001 *L'italiano nella varietà dei testi. L'incidenza della variazione diacronica, testuale e diafasica nell'annotazione e interrogazione di corpora generali e settoriali* – Coordinatore: Carla Marello.

[4] Nel prossimo futuro continuerò la mia ricerca avvalendomi anche dei corpora specialistici in italiano e in francese relativi alla fotografia (NUNC-IT Fotografia e NUNC-FR Fotografia) e dei rispettivi corpora generici in italiano e in francese (NUNC-IT Generico e NUNC-FR Generico), affinando le tecniche di ricerca.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 285-296.

I corpora che ho utilizzato e che utilizzerò per questa ricerca sono stati compilati con testi provenienti da newsgroup, i quali sono stati opportunamente trattati con speciali script per il loro specifico impiego[5].

2. L'ESTRAZIONE DEGLI ANGLISMI. I corpora italiani, cui ho fatto cenno nel paragrafo precedente, sono già stati utilizzati per analizzare più in generale l'impiego di forestierismi in italiano da parte degli utenti dei newsgroup (cfr. Valle 2006).

Nel corso della mia ricerca, ho preferito concentrare la mia attenzione verso gli anglismi in italiano ed in francese da un punto di vista sia interlinguistico che intralinguistico[6], quindi ho provato a ricercare gli anglismi mediante strategie maggiormente orientate verso questo obiettivo. L'utilizzo dei corpora NUNC si è rivelato utile, oltre che per lavori dal punto di vista terminologico, anche per lavori di carattere lessicografico, come ad esempio, l'integrazione del lemmario di un dizionario (cfr. Valle 2005 *i.s.*).

Le tecniche di ricerca sono ancora in fase di perfezionamento[7] per cui ora presento la metodologia adottata per questo lavoro. Gli anglismi sono stati estratti dai corpora specialistici italiani NUNC-IT Cucina e NUNC-IT Motori attraverso CQP (Corpus Query Processor, sviluppato presso l'IMS di Stoccarda)[8]. La metodologia adottata è di tipo corpus-based e mediante un apposito script, è stata effettuata una interrogazione in locale dei corpora, ottenendo così due liste POS nome, cioè due liste di nomi in ordine alfabetico con il loro numero di occorrenze presenti all'interno dei corpora citati. Da queste liste è stato possibile ricavare gli anglismi attraverso uno spoglio manuale delle liste stesse. Dopo la redazione di due liste contenenti gli anglismi presenti all'interno dei corpora di riferimento, sono stati ricavati i contesti in cui tali anglismi sono stati utilizzati. Questi contesti sono stati ottenuti interrogando i corpora online in cosiddetta "modalità linguistica".

Per tentare un confronto di tipo interlinguistico degli anglismi in francese ed in italiano (§ 4), ho effettuato la ricerca degli anglismi ricavati dai corpora italiani di riferimento, all'interno dei corpora specialistici francesi NUNC-FR Cucina e NUNC-FR Motori. Il reperimento dei dati dai corpora francesi è avvenuto interrogando direttamente i corpora online, con la modalità linguistica. In questo modo, ho ottenuto due liste di anglismi presenti all'interno dei corpora NUNC-FR Cucina e NUNC-FR Motori.

3. I RISULTATI OTTENUTI. Elenco qui di séguito gli anglismi ricavati interrogando i corpora di riferimento.

Gli anglismi estratti dal corpus NUNC-IT Cucina sono:

> *abuse, abstract, advisoring, account, agent, after hour, agribusiness, agrifood, angus beef, appetizer, apple pie, attack, baby sitter, baby-vegetables, background, bacon, backup, banner, bar, barbecue, barman, bean, bed&breakfast, beer, beerhunter, beerlander, beer lover, beerman, beer shop, beer taster, beer tasting, beginner, benchmark, biscuit, bitmap, black list, black bean, blend, blob, blueberry, body, bodybuilder, bodybuilding, boil time, book, bookcrossing, bookmark, boom, boomerang, ballot box, branch, bread, bread machine, breaking, brewer,*

[5] Per maggiori informazioni sulla preparazione dei corpora NUNC e sul trattamento dei testi confluiti al loro interno, rimando sempre agli articoli di Barbera ¶ 1 e Corino ¶ 13 cit., ed a Casavecchia 2005.

[6] Attualmente posso solo tentare un approccio di tipo interlinguistico (francese-italiano) dal momento che, per un confronto di tipo intralinguistico, devo ancora procedere con le ricerche degli anglismi all'interno dei corpora generici NUNC-IT Generico e NUNC-FR Generico. Successivamente potrò fare un confronto degli anglismi reperiti nel corpus generico della lingua di riferimento (in questo caso, italiano e francese) incrociando i dati ottenuti con quelli provenienti dai rispettivi corpora specialistici.

[7] Inizialmente avevo optato per una estrazione basata su gruppi consonantici significativi, cfr. Valle 2004, cit.; poi ho preferito adottare la metodologia che illustro in questo lavoro. Per il mio progetto di ricerca sviluppato in seno al Dottorato, queste tecniche potranno essere ulteriormente perfezionate.

[8] Per maggiori informazioni di carattere tecnico su CQP, rimando ad Heid ¶ 4, in questo volume.

brewhouse, brewmaster, brew-pub, brick, bricocenter, browser, brunch, budget, bug, bunker, business, businessman, buttermilk, buyer, byte, carawheat, catering, ceddar, cellophan, check-list, cherry, cheese cake, chinatown, chips, chipset, chutney, click, cocktail, coffee, company, compilation, cookie, copyright, cornflakes, corn sugar, cracker, crosspost, crosspostare, curry, database, design, director, directory, discopub, discount, dish, display, download, draft, draught, drink, drinker, drinking, dry-hopping, dummy, e-business, editing, editor, email, emoticon, entry level, establishment, export, factory outlet, fair play, fan, feedback, file, fitness, flag, flame, flavour, flop, floppy disk, folklore, footing, franchising, freezer, gadget, gazebo, grain, hacker, hall, hamburger, handball, handicap, happening, happy hour, hard disk, header subject, high gravity, highlands, hobby, homebrew, homebrewer, homebrewing, host, hostess, house, housing, iceberg, icewine, imprinting, improvement, input, internet, jogging, junk food, ketchup, keyword, killare, killer, killfile, killfilter, kit, knowhow, lady, lag phase, laptop, leader, leadership, link, linkare, linking, lobby, loss leader, low carb, low carber, lurkare, lurkaggio, lurkatina, lurkatura, lurker, lurking, mail, mailbox, mailer, mailing list, market, marketing, masher, mashing, master, meeting, megabyte, merchandising, morphing, new entry, netiquette, network, newbie, newsgroup, newsletter, newsreader, newsserver, nick, nickname, night, oatcake, optional, outsider, packaging, pancake, party, password, pastamaker, peanuts, pickles, pitch, pitching, please, plonk, plonkare, plug, plum cake, popserver, popup, post, postare, posting, powdery, privacy, private banker, problem, provider, pub, pusher, quotare, quoting, reader, reception, record, reply, rock, roast beef, rush hour, sandwich, scanner, scoop, scooter, screensaver, scripting, sherry, shop, shopper, shopping, shortbread, skylight, slang, slogan, smog, snack, software, spammer, spamming, spleen, sponsor, sponsorizzare, sponsorizzazione, springbank, stand, standard, stock, stress, stretching, subject, suffolk, supermarket, takeaway, thread, ticket, ticket restaurant, toast, toner, training, troll, trollare, trollata, trollazzo, trollismo, trub, vip, waffle, wafflemaker, watery, webcam, webmaster, weekend, welfare, whisky, whisky-brewer, winebar, winery, workshop.

Gli anglismi estratti dal corpus NUNC-IT Motori sono i seguenti:

abuse, accommodation barge, account, aquaplaning, adapter, advisor, aftermarket, agent, air-bag, airbox, anti-submarine, audience, autobus, automotive, baby, backbone, backgammon, background, backprotector, backstage, backup, badge, bancode, band, banner, bar, barman, beemer, bios, blacklist, blinker, blister, blockshaft, bloster, blowfish, board, bookmark, boom, boomerang, boost, booster, bounce, box, brainstorming, brake, brand, bull bar, briefing, broker, budget, bug, bunker, bus, business, business man, buyer, buzzer, bypass, byte, cab, call center, camera car, cameraman, car, car configurator, carshop, card, cash, cd charger, cello-phane, changer, chat, chattare, check, checklist, checkpoint, check panel, checkup, chipset, city car, client, clutch, cluster, cockpit, cocktail, comfort, common rail, common sphere, compact disc, computer, concept car, confort, cookie, cordless, country, crash, credit card, cross, cross-fire, crosspost, crosspostare, cummins, customer care, customer satisfaction, dealer, debugging, design, designer, desk, desktop, detector, dialer, direct, display, double-cab, download, drag-ster, driver, e-commerce, editing, e-mail, entry, entry level, facelift, facelifting, factory, fans, feedback, feeling, fiction, flame, flooding, flop, form, frame, franchising, free shop, full optional, gadget, gallery, gentlemen, glamour, go-kart, gossip, gps, grip, group, guard-rail, guest book, hacker, haldex, hall, handicap, handling, happening, hard disk, hardware, header, helper-spring, hobby, holding, homepage, interbusiness, intercooler, internet, instant book, jeans, jeep, joystick, jumbo, kart, keycard, killare, killer, killfile, know-how, layout, leader, leadership, leasing, link, linkare, lobby, loudness, lurkare, lurker, lurking, mail, mailbombing, mailbox, mailer, manager, marketing, master, meeting, metal detector, morphing, motorcycle, motor-home, multilink, naftonwagon, netiquette, network, newbie, new entry, news, newsgroup, news-letter, newsmaster, newsserver, nick, nickname, optional, outline, outlet, outsider, outsourcing, paddles, part time, pass, password, pickup, plonk, plonkare, plug, post, posting, postare, power boost, quoting, quotare, racing, redline, restyling, retrofit, road book, roll bar, safe boot, safety car, salesman, scooter, seatbacks, seatbelt, shock, shop, shopping, show car, showroom, side-bag, sidecar, skate board, silent-block, single-cab, slang, slogan, smartcard, snorkel, software, sound, sound blaster, spam, spammare, spammatore, spammer, spamming, speed control,

spoiler, sport, spray, sprint, states, status symbol, stock, stress, subject, sulky, sulphur, switch, test, testdrive, testdriver mode, tester, testing, thread, ticket, tool, top car, traction control, track challenge, track random, tracklist, trailer, training, trend, troll, trollata, trollare, trollaggio, trolleggiamento, trolleggiare, trolleggiata, trolleggiatore, trollone, trollonzo, truck, tuner, tuning, username, vintage, wagon, web, webcam, webagency, webmaster, website, yacht.

All'interno di questi è possibile rilevare la presenza di alcuni adattamenti:

chattare, crosspostare, killare, linkare, lurkare, lurkaggio, lurkatina, lurkatura, plonkare, postare, quotare, spammare, spammatore, sponsorizzare, sponsorizzazione, trollata, trollare, trollaggio, trolleggiamento, trolleggiare, trolleggiata, trolleggiatore, trollone, trollonzo.

È possibile notare che gli anglismi appartenenti all'àmbito cucina e latamente legati al cibo rappresentano un terzo del totale:

after hour, agribusiness, agrifood, angus beef, appetizer, apple pie, baby-vegetables, bacon, bar, barbecue, barman, bean, bed&breakfast, beer, beerhunter, beerlander, beer lover, beerman, beer shop, beer taster, beer tasting, biscuit, black bean, blend, blueberry, branch, bread, bread machine, breaking, brewer, brewhouse, brewmaster, brew-pub, brick, brunch, buttermilk, carawheat, catering, ceddar, cherry, cheese cake, chips, chutney, cocktail, coffee, cornflakes, corn sugar, curry, draught, drink, drinker, drinking, cracker, flavour, freezer, hamburger, happy hour, highlands, homebrew, homebrewer, homebrewing, icewine, junk food, ketchup, lag phase, low carb, low carber, masher, mashing, oatcake, pancake, party, peanuts, pickles, plum cake, roast beef, sandwich, sherry, shortbread, Suffolk, supermarket, takeaway, ticket, ticket restaurant, toast, trub, waffle, wafflemaker, watery whisky, whisky-brewer, winebar, winery.

Per stilare questa lista si sono rese necessarie ulteriori verifiche che disambiguassero eventuali termini polisemici. Si è controllato se "cookie", per esempio, occorresse nella sua potenziale valenza informatica, come accade effettivamente nei NUNC-Cucina italiani ([1a]), ma non nei NUNC-Cucina francesi ([1b]), dove è usato nel significato di "biscotto",

[1a] `Io uso Opera 6.5 che mi blocca tutti i` **`cookie`** `spioni , per cui mi avvisa con una finestra , a differenza di Explorer che li accetta automaticamente` NUNC-IT Cucina,

[1b] `en m' arr êtant net sur le trottoir et en regardant , consternée , mon` **`cookie`** `croquant à la cannelle rouler dans le caniveau crasseux` NUNC-FR Cucina,

insieme ad altre 4 occorrenze in contesti però di lingua inglese (ciò vale anche per i NUNC italiani di cucina; in questo tipo di testi inglesi si tratta sempre di veri e propri "biscotti" e non "cookies" informatici).

Anche il numero di anglismi che ha a che fare con le automobili, i motori e la loro vendita costituisce un terzo circa del totale degli anglismi presenti nei NUNC-Motori italiani:

barge, aquaplaning, adapter, airbag, airbox, anti-submarine, autobus, automotive, back-protector, beemer, blinker, blockshaft, bloster, boost, booster brake, brand, bull bar, camera car, cameraman, car, car configurator, carshop, check panel, city car, clutch, cockpit, comfort, common rail, common sphere, concept car, confort, customer care, customer satisfaction, dealer, dragster, driver, full optional, gadget, go-kart, gps, grip, guard-rail, haldex, helper-spring, intercooler, kart leasing, keycard, motorcycle, motorhome, naftonwagon, optional, paddles, pickup, power boost, racing, redline, restyling, retrofit, road book, roll bar, safe boot, safety car, salesman, scooter, seatbacks, seatbelt, shock, shop, shopping, show car, showroom, side-bag, sidecar, skate board, silent-block, single-cab, snorkel, speed control, spoiler, sport, spray, sprint, sulky, sulphur, switch, test, testdrive, testdriver, top car, traction control, truck, wagon.

Dalla verifica del numero di occorrenze, gli anglismi legati al medium ed alla comunicazione mediata dal computer rappresentano una sezione determinante del gruppo, costituendo quasi un terzo delle occorrenze (89) nella lista di NUNC-IT Cucina,

> *account, agent, backup, banner, bitmap, bookmark, browser, bug, byte, chipset, click, copyright, crosspost, crosspostare, database, directory, display, download, draft, e-business, editing, editor, email, emoticon, file, flag, flame, floppy disk, hacker, hard disk, header subject, input, internet, keyword, killare, killer, killfile, killfilter, laptop, link, linkare, linking, lurkare, lurkaggio, lurkatina, lurkatura, lurker, lurking, mail, mailbox, mailer, mailing list, megabyte, netiquette, network, newbie, newsgroup, newsletter, newsreader, newsserver, nick, nickname, password, plonk, plonkare, popserver, popup, post, postare, posting, provider, quotare, quoting, reader, reply, scanner, screensaver, scripting, software, spammer, spamming, subject, thread, toner, troll, trollare, trollata, trollazzo, trollismo, webcam, webmaster,*

e 92 nella lista di NUNC-IT Motori (ovvero, nuovamente, circa il 30% del totale),

> *account, adapter, agent, backup, banner, bookmark, byte, chat, chattare, chipset, client, compact disc, computer, cookie, cordless, crash, crosspost, crosspostare, debugging, desktop, display, download, e-commerce, editing, e-mail, flame, flooding, hacker, hard disk, hardware, header, homepage, internet, joystick, killare, killer, killfile, layout, link, linkare, lurkare, lurker, lurking, mail, mailbombing, mailbox, mailer, morphing, netiquette, newbie, newsgroup, newsletter, newsmaster, newsserver, nick, nickname, outsourcing, password, plonk, plonkare, plug, post, posting, postare, quoting, quotare, smartcard, software, sound blaster, spam, spammare, spammatore, spammer, spamming, subject, thread, troll, trollata, trollare, trollaggio, trolleggiamento, trolleggiare, trolleggiata, trolleggiatore, trollone, trollonzo, username, web, webcam, webagency, webmaster, website.*

come, per esempio, in:

[2] `Con la presente si segnala il` **`flooding`** `effettuato nei confronti del newsgroup da parte di un vostro utente , inviando numerosi messaggi da centinaia di kb ciascuno sul newsgroup di cui sopra` NUNC-IT Motori.

Alcuni termini possono tuttavia ricorrere sia con un significato legato alla sfera semantica dei motori, sia con un significato tipico della CMC, cfr. *crash* (su 103 occorrenze riscontrate, 64 si presentano nell'espressione fissa "crash test"; il secondo esempio appartiene invece tipicamente al linguaggio informatico),

[3a] `comunque la punto ha 4 stelle nei` **`crash`** `test e si comporta al meglio al pari di lupo e polo nella sua categoria` NUNC-IT Motori,

[3b] `ebbene sì , stamattina mi è andato in` **`crash`** `il navigatore gps della bmw .. cercavo una strada e si è bloccato tutto il computer di bordo` NUNC-IT Motori.,

e *morphing* (notare, tra l'altro, come il terzo esempio giochi con il termine):

[4a] `Pratichi anche un bel` **`morphing`** `del tuo bel nome per evitare di essere filtrato . Complimenti . Tanto da quell' " audirull " e dal " ke " invece di " che " si vede subito che sei proprio ... il tuo nome non lo faccio , che è meglio ...` NUNC-IT Motori,

[4b] `Ah quello è` **`morphing`** `? io pensavo che fare morphing era cambiare nick per non farsi riconoscere !!` NUNC-IT Motori,

[4c] `Ciao a tutti , preso da impeto di changing (` **`morphing`** `, alias tamarring ), vorrei montare dei cerchi da 15" sulla mia Bravo ...` NUNC-IT Motori,

Per quanto riguarda i dati di newsgroup francesi, gli anglismi presenti nel NUNC-IT Cucina ed estratti anche dal corrispondente NUNC-FR Cucina sono i seguenti:

after-dinner, after-shave, apple crisp, background, bacon, banana bread, bar, barbecue, barman, bean, beer, beer agency, beer engines, beer journaliste, beershop, black pudding, body builder, boomerang, break, brewferm, brewpub, brewmaster, browser, brunch, budget, bug, bunker, business, buttermilk, bookmark, catering, cherry pie, cheese cake, chips, chutney, cocktail, compilation, cookie, copyright, cornflakes, cracker, crosspost, curry, design, discount, drink, drink market, dry hopping, e-business, email, e-mail, emoticon, fair play, fan, feedback, flag, flop, folklore, freezer, gadget, hall, hamburger, handicap, hard discount, highland, hobby, iceberg, icewine, instant check, internet, ketchup, kill-file, killfile, killer, kit, lady, leader, link, lobby, long drink, lurkage, lurker, lurkeur, mail, mailing list, marketing, master, meeting, netiquette, newbie, newsgroup, newsreader, nick, packaging, pancake, party, peanuts, pickles, plonk, plonker, popup, post, poster, posting, provider, pub, reception, record, rock, roast-beef, roastbeef, sandwich, scoop, cherry, shop, shopping, shortbread, slang, slogan, snack, software, spam, spammer, sponsor, standard, stock, stress, supermarket, thread, ticket, toast, troll, webcam, webmaster, weekend, whisky, white pudding, winery, workshop.

Gli anglismi presenti nel NUNC-IT Motori ed estratti anche dal corpus NUNC-FR Motori sono:

aquaplaning, aftermarket, airbag, audience, autobus, background, backstage, backup, badge, barman, bios, blister, bookmark, boom, boomerang, boost, booster, box, briefing, broker, bug, bunker, bus, business, buzzer, bypass, cameraman, car pass, cd, cellophane, check-up, checklist, cocktail, comfort, common rail, confort, cookie, crash test, cross, crosspost, dealer, design, designer, desktop, dragster, driver, e-mail, email, fan, feedback, feeling, fiction, flooding, flop, frame, full option, gadget, glamour, gps, hall, handicap, hobby, holding, homepage, intercooler, internet, jeans, jeep, joystick, jumbo, kart, keycard, killfile, kill-file, leader, leasing, link, lobby, loudness, lurkage, lurker, lurkeur, lurkeuse, mail, mailbox, marketing, master, meeting, netiquette, newbie, news, newsgroup, newsletter, nick, nickname, outsider, pass, pickup, plonker, plug, post, racing, restyling, safety car, serial killer, scooter, shop, shopping, show room, sidecar, skate, silent bloc, slogan, software, sound system, spam, spamming, spammer, spammeur, spoiler, sport, spray, sprint, stock, stress, switch, test, thread, ticket, trend, troll, troller, tuner, tuning, vintage, wagon, web, webcam, webmaster, yacht.

All'interno di questi ci sono solo *lurkage*, *lurkeur*, *lurkeuse* e *spammeur* come adattamenti alla morfologia francese.

4. PRIMO APPROCCIO INTERLINGUISTICO TRA ANGLISMI NEI CORPORA ITALIANI ED ANGLISMI NEI CORPORA FRANCESI. Durante la ricerca, è emerso l'utilizzo da parte degli utenti italiani di *lurker* che in italiano vale 'utente di un newsgroup che legge i messaggi, senza partecipare al dibattito mediante l'invio di risposte ai messaggi letti'. Qui di séguito fornisco alcuni esempi:

[5a] Tu , avendo fatto outing , non hai più diritto allo status di **lurker** . NUNC-IT Cucina,

[5b] Infatti la presenza di **lurkers** qui fuori non può essere provata perché il **lurker** è ignoto per definizione , nel momento in cui si palesa non è più tale e quindi l' outing di un **lurker** non costituisce prova dell' esistenza degli stessi . NUNC-IT Cucina.

Questo anglismo è molto usato dagli utenti dei newsgroup, anche senza adattamenti alla lingua italiana (come, oltre al maschile singolare, il maschile plurale *lurkers* utilizzato nell'esempio [5b]), insieme al prestito adattato con morfologia italiana derivativa *lurkatore* che presenta lo stesso significato di *lurker*. Per esempio:

[6a] nessuno è tenuto a rispondermi , a maggior ragione visto che non mi avete neanche mai visto (sono un **lurkatore**) ... cmq ... Io AMO mangiare e soprattutto mangiare bene NUNC-IT Cucina,

[6b] Dopo una vita da **lurkatore** prendo il coraggio e scrivo il mio primo intervento , pardon il secondo . NUNC-IT Cucina.

Anche alcuni utenti dei newsgroup francesi utilizzano il prestito *lurker* senza adattarlo alle regole derivative della lingua francese (come, oltre al maschile singolare, il maschile plurale *lurkers* nell’esempio [7a]), mentre altri utenti utilizzano il prestito adattato con morfologia derivativa *lurkeur* (anche al maschile plurale *lurkeurs*). Il significato di entrambi gli anglismi è lo stesso di *lurker* e *lurkatore* utilizzato dagli utenti italiani, ed i due prestiti in francese, *lurker* (ess. [7]) e *lurkeur* (ess. [8]), coesistono al pari di quelli italiani. Qui di sèguito fornisco alcuni esempi per quanto riguarda il francese:

[7a] Un grand bravo donc à tous les lecteurs de frbv , contributeurs notoires ou **lurkers** anonymes , qui ont été brillamment reçus au TGQE . A la saison prochaine , Philippe Steff Bonne vacances Corinne NUNC-FR Cucina,

[7b] lassant trop vite du banal flanc patissier Me voilà , finalement , prêt à me lancer Et à passer de **lurker** à contributeur . Rougisseant à l' avance d ' être par trop banal Décidé à exister dans ce lieu cordial NUNC-FR Cucina;

[8a] Ils ne participent pas tous autant , mais ils y sont . Et je ne compte pas les **lurkeurs** ... Et ce n' est pas une question de leçon NUNC-FR Cucina,

[8b] salut Christian et bonne année ainsi qu' aux zaut' , contributeurs ou **lurkeurs** ..." Christian Callec " . NUNC-FR Cucina,

[8c] Un **lurkeur** qui se met à poster , il perd ipso facto sa qualité de **lurkeur** NUNC-FR Generic I.

Si evince dai contesti che una serie di forme *lurker* occorre in francese come forma infinita del verbo (cfr anche es. 12b; tendenza inoltre confermata dai risultati dei NUNC-FR generici), solo una parte è costituita dal sostantivo inglese. Possiamo pertanto ipotizzare che proprio la necessità di disambiguazione abbia reso necessaria la diffusione della forma adattata *lurkeur* e la compresenza dei due sinonimi, di cui ritroviamo anche il corrispondente femminile *lurkeuse*:

[9] Tout-à-fait indépendemment de frc et en tant que **lurkeuse** de fufe je trouve que , au même titre que tu dis (et d' autres aussi) que tout le monde à son mot à dire sur créations , changements , destructions de forums il faut veiller à ce que ce soit * vraiment * le cas , C' est le cas . NUNC-FR Generic II.

Per rimanere nello stesso àmbito, gli utenti italiani utilizzano molto il verbo *lurkare* nel valore di ‘leggere i messaggi di un newsgroup senza partecipare al dibattito rispondendo ai messaggi letti’. Di questo prestito adattato sono stati trovati numerosi contesti:

[10a] prima di iniziare a postare è necessario **lurkare** IDA per un po' di tempo ; NUNC-IT Motori,

[10b] a furia di stare a **lurkare** (e nell' occasione a scrivere) il lavoro d' ufficio non va avanti e rischio il licenziamento ! NUNC-IT Motori,

[10c] Non scrivo molto sul NG siccome mi piace più **lurkare** , però stavolta non posso esimermi di raccontarvi la mia " avventura " . NUNC-IT Motori.

Inoltre gli utenti utilizzano il verbo *lurkare* con relative coniugazioni; ad esempio:

[11a] E dato che **lurko** da moooolto tempo , senza mai intervenire nei discorsi , e notando gente moooolto preparata in materia birrofila e visto che la birra x me è una religione
NUNC-IT Cucina,

[11b] Salve a tutti , sono una ragazza di Napoli e **lurko** già da un po' su questo ng , e mi sono decisa a scrivere perchè c' è una disputa nella mia famiglia NUNC-IT Cucina,

[11c] Direi che siamo nella fascia e nel genere di locali tipo il Savoia (che tanto piace all' amica mafe che ogni tanto **lurka** e interviene qui su idr) . NUNC-IT Cucina,

[11d] Abbiamo speso attorno ai 30-35 euro a testa (eravamo in cinque , e almeno altri due **lurkano** di tanto in tanto IDR) comprese due bottiglie passabili , l' ambiente è semplice ma carino e la cameriera aveva un sorriso che faceva innamorare :-)
NUNC-IT Cucina,

[11e] Salve a tutti , vi ho **lurkato** giusto giusto stasera e penso che possiate darmi qualche consiglio se vi va . NUNC-IT Cucina,

[11f] Dopo aver **lurkato** a lungo , inizio a dare anch' io un contributo al newsgroup nel modo, spero, piu' apprezzato, cioe' con una recensione ... NUNC-IT Cucina.

Anche gli utenti francesi, come già accennato, utilizzano il verbo *lurker*, con lo stesso significato del verbo italiano *lurkare*, anche se sembrano più restii a coniugarlo. Infatti sono stati trovati pochi contesti. Eccone un paio:

[12a] Salut Patrick , Tu vois , je **lurke** encore . Bien le bonjour à vous trois de nous trois (Fabienne , Thibault et moi-même) .
NUNC-FR Cucina,

[12b] j' ai d' autre préocupations ... je vous dit pas adieu ni au revoir , je continue à **lurker** frm)) voilà ... NUNC-FR Motori.

Inoltre, all'interno dei newsgroup italiani e francesi, gli utenti utilizzano molto i prestiti adattati con morfologia derivativa *lurkaggio* (italiani, ess. [13]) e *lurkage* (francesi, ess. [14]), entrambi col significato di 'attività di lettura dei messaggi di un newsgroup senza partecipazione al dibattito mediante la risposta ai messaggi letti'. Fornisco qui di séguito alcuni esempi:

[13a] Dopo mesi e mesi di **lurkaggio** e sfruttamento dei suggerimenti del NG, penso sia giunta l' ora di sdebitarmi . NUNC-IT Cucina,

[13b] allora , visto e considerato che dopo tanto **lurkaggio** ho iniziato a postare mi sembra giusto contribuire con una ricetta , ovviamente sarda , e di facile preparazione NUNC-IT Cucina;

[14a] ou de lecture peu ou prou attentive pendant une durée raisonnable (fut-un temps où on conseillait à semaines de **lurkage** avant de commencer à poster , tout se perd mon bon monsieur) t' aurais montré que le forum NUNC-FR Cucina,

[14b] Bonjour à tous , ceci est mon premier post sur ce newsgroup qui après quelque temps de **lurkage** m' a déjà permis d' apprendre pas mal de choses) Voici donc mon problème . NUNC-FR Motori.

Dagli esempi si può dedurre la produttività del prestito inglese *lurker* attestato nel *Longman* monolingue inglese (LDCE) con il significato con cui è usato sia in italiano, sia in francese: 'if you lurk in a chat room in the Internet, you read what other people are writing to each other, but you do not write any messages yourself'.

L'italiano, più ricco di morfologia alterativa, vi ricorre come mezzo espressivo (cfr. anche Dressler - Merlini Barbaresi 1994), presentando un più ampio ventaglio di varianti scherzose degli adattamenti, ad es. *lurkatina* e *trollonzo*. La caratteristica "giocosa" della lingua dei newsgroup si esprime anche in questo modo oltre che con gli emoticon e l'uso di caratteri maiuscoli o lettere ripetute (Gheno 2004). Si veda in proposito anche l'uso di scorciamenti studiati da Allora - Marello *i.p.*

4.1 TRA *VOYEUR* E *LURKER*. Nell'àmbito delle rigide normative ministeriali circa l'uso dei forestierismi nella lingua francese, la frequenza di *lurker* è emblematica, visto che il francese ha a sua disposizione l'internazionalmente noto *voyeur*.

Alcuni contesti nei newsgroup francesi mostrano effetttivamente casi di *voyeur* con lo stesso significato di *lurker* (cfr. ess. [15]), a volte con esplicito passaggio metaforico (es. [16]), con riferimento alla sfera uditiva (es. [17]), o con consapevole riflessione metalinguistica (es. [18]):

[15a] Dommage vraiment que toute cette faune de **voyeurs** ne participe pas un peu de temps en temps ici ! Putain faut oser c' est pas bien compliqué de poster. NUNC-FR Foto,

[15b] (**voyeur**) désolé ^^ , je vais faire un effort) je sens que je vais bien apprécier ce newsgroup , ça a l air asser marrant) héhé NUNC-FR Foto,

[15c] je retourne à mon mutisme de **voyeur** avide PS : Tartineau , arrête de nourrir les trolls) NUNC-FR Generic I;

[16] Ne pas le faire par mail perso , histoire de . Car , cela en est gênant de vous lire , j' ai eu l' impression d' être un **voyeur** , et de suivre une conversation privée . NUNC-FR Generic I;

[17] Les communications de Sarko sont mieux protégées moins analogiques et font tomber sous les foudres de la loi le super malin plus informaticien et moins bricoleur mais toujours **voyeur** des oreilles ... C' est complètement illégal de raconter ce que l' on a écouté et même de donner la fréquence ... NUNC-FR Generic I;

[18] C' est donc dire que ce n' est qu' apr ès un certain temps qu' il est possible de différencier les deux , puisqu' on ne peut savoir avec certitude si un nouvel arrivant sur un forum sera un apprenti ou un reluqueur . Mais le mot " **voyeur** " ne ne pas remplacer efficacement " reluqueur " ? Sur ce , à bientôt , je m' en vais manifester ... NUNC-FR Generic I.

5. CONCLUSIONI. Come già dimostrato anche in altri lavori inclusi in questo volume, l'utilizzo dei corpora NUNC appare molto utile in diversi àmbiti. Per la mia ricerca sugli anglismi in francese ed in italiano, in particolare, i NUNC sono specialmente utili, poiché mostrano uno scritto non sorvegliato e, nel caso dei NUNC francesi, anche uno scritto eccezionalmente non influenzato dalle direttive ministeriali sul rifiuto dei prestiti inglesi.

Il passo successivo sarà l'analisi del genere grammaticale attribuito nelle due lingue ai prestiti inglesi e l'esame della grafia delle parole inglesi in italiano ed in francese. Inoltre si allargherà il contesto esaminando le collocazioni più significative degli anglismi in entrambe le lingue. Dal punto di vista statistico, sarà rilevante verificare se la percentuale relativa all'incidenza degli anglismi all'interno dei corpora generici e specialistici di entrambe le lingue sia la medesima o meno. Questo dato dovrebbe, tra l'altro, aiutare a chiarire se le politiche linguistiche adottate in Francia influenzino in maniera rilevante l'impiego degli anglismi da parte degli utenti: e le prime risultanze presentate in questo contributo farebbero propendere per il no.

BIBLIOGRAFIA.

ALLORA - MARELLO
i.p. Adriano Allora - Carla Marello, *"Ricarica clima". Accorciamenti nella lingua dei newsgroup*, contributo per il *IX congresso internazionale della Società di linguistica e filologia italiana (SILFI). Prospettive nello studio del lessico italiano, Firenze 14-17 giugno 2006*, in corso di stampa negli *Atti*.

BARBERA
2003 Manuel Barbera, *Review Article of "M. Görlach (ed.), A Dictionary of European Anglicisms"* in "International Journal of Lexicography" XII (2003)[2] 208-216.
2004 *in.* Manuel Barbera, *Il progetto FIRB. Stato dei lavori*, documento interno inedito, Ver. 7 aggiornata al febbraio 2004.
¶ 1 Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in questo volume, pp. 3-20.

BOWKER - PEARSON
2002 Lynne Bowker - Jennifer Pearson, *Working with Specialized Language. A Practical Guide to Using Corpora*, London - New York, Routledge, 2002.

CASAVECCHIA
2005 Sara Casavecchia, *Progettazione ed implementazione di corpora di lingua inglese basati sui newsgroup,* Università di Torino, Facoltà di Lingue, Tesi di Laurea, 2004-2005.

CHAURAND
1999 Jacques Chaurand, *Nouvelle histoire de la langue française*, Paris, Seuil, 1999.

COLIN
2003 Jean-Paul Colin, *Le lexique*, in YAGUELLO 2003, pp. 391-456.

CORINO
¶ 13 Elisa Corino, *NUNC est disputandum. Aspetti della testualità e questioni metodologiche*, in questo volume, pp. 225-252.

CORRÉARD
2002 *Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins*, edited by Marie-Hélène Corréard, s.l. (U.K.), EURALEX, 2002.

DEA→ GÖRLACH 2001.

DRESSLER - MERLINI BARBARESI
1994 Wolfgang U. Dressler - Lavinia Merlini Barbaresi, *Morphopragmatics. Diminutives and Intensifiers in Italian, German, and Other Languages,* Berlin - New York, Mouton de Gruyter, 1994.

GHENO
2004 Vera Gheno, *Prime osservazioni sulla grammatica dei gruppi di discussione telematici di lingua italiana*, in "Studi di Grammatica Italiana" XXII (2004) 267-308.

GÖRLACH
2001 *A Dictionary of European Anglicisms: A Usage Dictionary of Anglicisms in Sixteen European Languages*, edited by Manfred Görlach, Oxford, Oxford University Press, 2001.
2002 *An annotated bibliography of European Anglicisms*, edited by Manfred Görlach, Oxford, Oxford University Press, 2002.

GREFENSTETTE - NIOCHE
2000 Gregory Grefenstette - Julien Nioche, *Estimation of English and non-English Language Use on the WWW*, in *Proceedings of RIAO* [Recherche d'Informations Assistée par Ordinateur] *2000, "Content-Based Multimedia Information Access", Paris, 12.-14. April 2000*, Paris, 2000, pp. 237-246, disponibile online alla pagina `http://arxiv.org/pdf/cs.CL/0006032`.

GREFENSTETTE
2002 Gregory Grefenstette, *The WWW as a Resource for Lexicography*, in CORRÉARD 2002, pp. 199-215.

GROSSMANN - RAINER
2004 *La formazione delle parole in italiano*, a cura di Maria Grossmann e Franz Reiner, Tübingen, Max Niemeyer Verlag, 2004.

GUSMANI
1986 Roberto Gusmani, *Saggi sull'interferenza linguistica*, Firenze, Le Lettere, 1986.
1989 Roberto Gusmani, *Interlinguistica*, in LAZZERONI 1989, pp. 87-114.

HEID
¶ 4 Ulrich Heid, *Il* Corpus WorkBench *come strumento per la linguistica dei corpora. Principi ed applicazioni*, in questo volume, pp. 89-108.

KILGARRIFF - GREFENSTETTE
2003 Adam Kilgarriff - Gregory Grefenstette, *Introduction to the Special Issue on the Web as Corpus*, in "Computational Linguistics" XXIX (2003)[3] 333-347, disponibile anche online alla pagina `http://www.kilgarriff.co.uk/ publications.htm`.

LAZZERONI
1989 *Linguistica storica*, a cura di Romano Lazzeroni, Roma, Carocci, 1989.

LDCE
2003 *Longman Dictionary of Contemporary English, New ed.*, Harlow, Longman, 2003.

LÓPEZ DÍAZ - MONTES LÓPEZ
2006 *Perspectives fonctionnelles: emprunts, économie et variations dans les langues. S.I.L.F. 2004. XXVIII Colloque de la Société internationale de linguistique fonctionnelle, tenu à Saint-Jacque-de-Compostelle et à Lugo du 20 au 26 septembre 2004*, édité par Motserrat López Díaz et María Montes López, Lugo, Editorial Axac, 2006.

MARELLO
1996 Carla Marello, *Le parole dell'italiano. Lessico e dizionari,* Bologna, Zanichelli, 1996.

NODE
2000 *The New Oxford Dictionary of English on CD-ROM*, Oxford, Oxford University Press, 2000.

REY-DEBOVE - GAGNON
1990 Josette Rey-Debove - Gilberte Gagnon, *Dictionnaire des anglicismes*, Paris, Le Robert, 1990.

VALLE
2005 *i.s.* Luca Valle, *The Retrieval of Anglicisms in Newsgroups Usenet Corpora (NUNC)*, comunicazione a *JILC 2005. 4èmes Journées Internationales de la linguistique de corpus, Lorient 15-17 septembre 2005*, Université de Bretagne Sud, 2005, in corso di stampa.

2006 Luca Valle, *Varietà diafasiche e forestierismi nell'italiano nei gruppi di discussione in rete*, in LÓPEZ DÍAZ - MONTES LÓPEZ 2006, pp. 371-374.

YAGUELLO
2003 *Le grand livre de la langue française*, sous la direction de Marina Yaguello, Paris, Seuil, 2003.

CORPORA E SITI DI RIFERIMENTO.

corpora.unito.it	`http://www.corpora.unito.it/`.
NUNC	`http://www.bmanuel.org/projects/ng-HOME.html`.
NUNC-IT Cooking	`http://www.bmanuel.org/projects/ng-HOME.html`.
NUNC-IT Motor	`http://www.bmanuel.org/projects/ng-HOME.html`.
NUNC-FR Cooking	`http://www.bmanuel.org/projects/ng-HOME.html`.
NUNC-FR Motor	`http://www.bmanuel.org/projects/ng-HOME.html`.

FELISA BERMEJO
(Università di Torino)

17. *Consigliare* / *aconsejar* e le subordinate esplicite od implicite.
Analisi contrastiva nei NUNC generici.

0. INTRODUZIONE. I corpora in cui è stata realizzata la ricerca sono il NUNC-IT generico di lingua italiana (I parte) ed il NUNC-ES generico di lingua spagnola di corpora.unito.it[1].

Il presente lavoro affronta lo studio delle subordinate rette dal verbo *consigliare* in italiano e dal verbo *aconsejar* in spagnolo, attraverso l'analisi delle occorrenze riscontrate nei corpora NUNC generici. Tale analisi permette di valutare la possibilità e la frequenza dell'uso delle subordinate implicite od esplicite dipendenti da questi verbi, partendo dal presupposto che nelle due lingue sono possibili entrambe le strutture.

Per l'italiano, un'importante questione da affrontare riguarda la presenza / assenza del complementatore preposizionale *di*. D'altra parte, esaminare il corpus permetterà di verificare se vi è qualche caso di *consigliare* con subordinata esplicita, considerato che la grammatica lascia aperta la possibilità di tale costruzione. Per lo spagnolo, del verbo *aconsejar* si esaminerà nel corpus, innanzitutto, la frequenza con cui viene usata la subordinata esplicita rispetto all'implicita; in secondo luogo, si verificherà se esistono contesti sintattici che possano determinare l'impiego dell'infinito nella subordinata, per poter così stabilire le condizioni d'uso di esplicita / implicita. Infine saranno esposte le simmetrie e dissimmetrie sintattiche dei due verbi.

Una volta delimitato l'uso di *consigliare* ed *aconsejar* come verbi di influenza[2], va fatta un'altra precisazione che riguarda il soggetto espresso della subordinata. Siccome in questo caso la subordinata è sempre esplicita, essa rimane fuori del nostro campo di analisi, che si concentra invece sulla possibile alternanza implicita / esplicita. Nei corpora NUNC, inoltre, si è riscontrata una sola occorrenza di ognuno dei verbi[3]; pertanto anche l'esiguo numero dei loro casi giustifica l'esclusione.

1. *CONSIGLIARE* ED *ACONSEJAR*: VERBI DI INFLUENZA SENZA SOGGETTO ESPRESSO NELLA SUBORDINATA. Nell'insieme dei verbi volitivi si può individuare un gruppo denominato verbi di influenza. Sono *pedir* ("chiedere"), *rogar* ("pregare"), *mandar*, *ordenar* ("comandare", "ordinare"), *permitir* ("permettere"), *prohibir* ("proibire", "vietare"), *aconsejar* ("consigliare"). I verbi di influenza presentano alcune caratteristiche semantiche e sintattiche comuni, come il fatto di reggere la subordinata con verbo al congiuntivo, se è esplicita, o con verbo all'infinito, se è implicita.

---

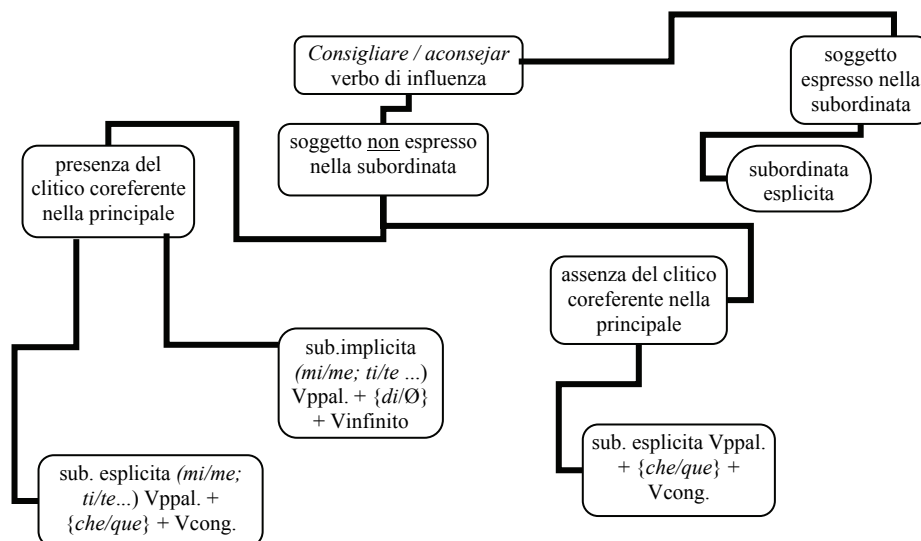
[1] Salvo diversamente indicato, tutti gli esempi in Courier devono intendersi tratti da questi due corpora. Naturalmente, tutti gli esempi sono stati riportati con la stessa ortografia e punteggiatura dell'originale.
[2] È doveroso far presente che i verbi *consigliare* ed *aconsejar* sono usati anche come verbi dichiarativi. Questo secondo impiego non è tuttavia oggetto di studio nel presente lavoro e nei corpora analizzati se ne è trovato un numero scarso di casi.
[3] Si vedano i due esempi seguenti:

[1a] `L' Organizzazione Mondiale della Sanità e l' UNICEF` **`consigliano che`** `i neonati siano alimentati esclusivamente con latte materno -- nient' altro , nemmeno acqua -- per i primi sei mesi circa`

[1b] `por riesgos de malformaciones congenitas se` **`aconseja que`** `las mujeres embarazadas no se tomen radiografias ni scanners.`

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 297-308.

Dai dati che emergono, l'interesse dell'analisi delle strutture di cui fanno parte *consigliare* ed *aconsejar* si riferisce ad aspetti diversi. Nel caso di *consigliare*, l'elemento rilevante consiste nell'assenza del complementatore preposizionale *di* in costruzioni copulative; nel caso di *aconsejar*, nella scelta della struttura esplicita od implicita. In stretto collegamento con tale scelta, gioca un ruolo determinante la presenza od assenza di un coreferente (clitico) nella principale.

1.1 ESPLICITA OD IMPLICITA? I termini della questione possono essere rappresentati con la tavola seguente:

Tav. 1: Subordinate esplicite ed implicite.

1.1.1 *CONSIGLIARE*. In italiano, il verbo *consigliare* si costruisce con subordinata implicita all'infinito. Sull'uso dell'esplicita, Renzi segnala che «la forma temporalizzata, per quanto non esclusa del tutto, è meno usuale e di livello stilisticamente piuttosto alto; inoltre, essa è spesso limitata a quei casi in cui il soggetto della subordinata può essere interpretato come impersonale. Da qui la frequenza, in questi costrutti, della costruzione con il *si* o della forma passiva». (*GGIC* II, p. 644).

1.1.2 *ACONSEJAR*. La scelta del parlante tra l'uso dell'esplicita o dell'implicita non risponde a fattori di indole sociolinguistica; si direbbe che l'uso dell'una o dell'altra forma sia indistinto. Di fatto le grammatiche, nell'affrontare la frase complessa ed analizzare nello specifico i verbi di influenza, non danno nessuna indicazione in proposito.

L'unico commento che è stato possibile reperire al riguardo è quello della studiosa Torrente Sánchez-Guisande (1998, p. 76), nel suo libro dedicato allo studio delle subordinate sostantive spagnole. In esso, in merito al verbo *aconsejar*, l'autrice segnala che la combinazione con subordinata implicita è poco frequente e non è raccomandabile. Tale rilievo è uno dei motivi per cui abbiamo iniziato l'esplorazione dei corpora.unito.it.

Un'altra motivazione è costituita da una "regla práctica" diffusa tra gli insegnanti di spagnolo, che ha l'obiettivo didattico di evitare interferenze e facilitare l'apprendimento delle strutture spagnole da parte degli italofoni. Nell'insegnamento delle subordinate spagnole a discenti di lingua italiana, si preferisce dare una regola generale (rispettata da tutto il gruppo di verbi di influenza) secondo cui la subordinata dipendente da tali verbi è sempre esplicita con il verbo al

congiuntivo. La scelta di insistere su questa regola, tralasciando la possibilità dell'implicita, è motivata da due importanti ragioni: (1) la costruzione è diversa da quella usata in italiano, dove si impiega l'implicita; (2) non tutti i verbi si possono costruire con l'implicita, che non è ammessa con verbi come *pedir*, *rogar* ed anche *decir*, quando viene usato come verbo d'influenza.

1.2 *CONSIGLIARE*. IMPLICITE CON E SENZA INTRODUTTORE *DI*. In base alle occorrenze tratte dai corpora, *consigliare* presenta diverse strutture.

1.2.1 *CONSIGLIARE* + *DI* + INFINITO. Tre i sottocasi da prendere in considerazione.

(a) *Consigliare* in voce attiva si attua in una struttura implicita introdotta da *di*: *consigliare* + *di* + infinito:

[2a] Prima lezione di chitarra elettrica , mi **consiglia di** impugnare il plettro tra indice ed anulare , tenendo diritto il pollice e le altre dita a pugno.

[2b] La bimba va dalla mamma a mostrarle il suo guadagno e la mamma si mostra ammirata e le **consiglia di** portarlo alla banca.

(b) *Consigliare* con *si* impersonale regge anche una subordinata infinitiva introdotta da *di*: *si consiglia* + *di* + infinito:

[3a] Se non si riceve nulla entro ore , pertanto **, si consiglia di** scrivere a cfv per avere informazioni .

[3b] Il disco del newsserver è pieno **: si consiglia di** riprovare l' invio dopo un po' di tempo .

(c) *Consigliare* in voce passiva perifrastica. L'infinito è preceduto da *di,* come alla voce attiva. Le occorrenze riscontrate sono al passato prossimo (*è stato* + *consigliato* + *di* + infinito):

[4a] Mi hanno detto che la preparazione alle varie università x fisica è praticamente la medesima , mi **è stato consigliato di** andare dove ci sono pochi studenti

[4b] ... alle nostre domande su come raggiungere Pisa ci **è stato consigliato di** prendere un normale autobus di linea .

1.2.2 *È CONSIGLIATO* + INFINITO. Nella ricerca delle forme verbali di *consigliare* abbiamo avuto occasione di trovare il verbo *consigliare* usato nella struttura: *è consigliato* + infinito, che si contraddistingue per il fatto che l'infinito non è introdotto dal complementatore preposizionale *di*. Si tratta, secondo Renzi (*GGIC* II, p. 673), di una frase copulativa, giacché nelle costruzioni copulative, a differenza delle passive, «l'introduttore *di* non appare»[4].

Nel corpus abbiamo trovato (tra i 1000 *matches* di *consigliato*) 18 occorrenze di frasi copulative con il participio *consigliato* nelle quali non appare l'introduttore *di*, cioè *è* + *consigliato* + infinito (l'infinito **non** è preceduto da *di*):

[5a] Peraltro **è consigliato** proteggere la terra con un foglio di plastica quando si spruzzano fitofarmaci .

[5b] Con questo tipo di lampada **è consigliato** lasciare la tarta per - 8 ore sotto la luce . E ' importante lasciare le tartarughe sotto una lampada

Il corpus analizzato ci ha offerto, però, un esempio con l'introduttore *di*, *è* + *consigliato* + *di* + infinito:

[4] Per un analisi più approfondita, Renzi rimanda alle strutture copulative predicative con un aggettivo in cui la subordinata funge da soggetto e non è preceduta da *di* (*GGIC* II, p. 661).

[6] `questa operazione è decisamente CPU intensive e quindi in caso di configurazione poco potente` **`è consigliato di`** `aggiungere i file in più riprese .`

Come dobbiamo considerare tale struttura, come passiva o come copulativa? Si direbbe che si tratti di una frase copulativa in cui l'inserimento del complementatore *di* è indice di un registro popolare.[5] Inoltre, si è trovata un'occorrenza con subordinata esplicita, *è* + *consigliato* + *che* + *verbo congiuntivo*:

[7] **`E'`** `stravivamente` **`consigliato che`** `tu prenda prima la MB e dopo ci metti su i nuovi componenti ...`

Questo è in assoluto l'unico caso con esplicita che si è trovato nel Corpus generico NUNC Italiano I. Anche questo esempio ci fa pensare ad una costruzione copulativa. Possiamo stabilire dei parallelismi con altri verbi di influenza come *vietare*, che è usato in frasi copulative modificato da un avverbio: *è severamente vietato fumare*. Dobbiamo anche ricordare che in verbi che esprimono l'idea di formazione o composizione quali *formare*, *costituire*, *comporre*, ecc., il participio è usato con valore aggettivale in frasi copulative: *Il mazzo di fiori è formato di/da 7 rose*. La distinzione è marcata differenziandone l'accezione nel dizionario Sabatini-Coletti (*DISC*).

1.3 *ACONSEJAR*. ESPLICITE ED IMPLICITE. Se raccogliamo i dati dei corpora in una tabella (cfr. Tav. 2 qui sotto), si osserva che, fatta eccezione per le forme *aconseja* ed *aconsejan*, il numero di occorrenze esplicite è uguale o superiore al numero di quelle implicite. In questo senso spicca il caso di *aconsejo*, con un numero di occorrenze esplicite assai superiore al doppio. Invece, *aconseja* presenta in tutte le occorrenze (tranne una) la subordinata implicita; va sottolineata anche la forma *aconsejan*, poiché non presenta nessun esempio di esplicite.

Forma verbale	**Esplicite (nº occ.)**	**Implicite (nº occ.)**	**Totale**
aconsejo	45	19	64
aconseja	01	12	13
aconsejamos	03	02	05
aconsejan	00	06	06
ha aconsejado	01	01	02
aconsejaba	01	01	02
aconsejé	02	01	03
aconsejó	04	03	07
aconsejaron	03	02	05
aconsejaría	05	02	07
aconsejaré	00	01	01
Totale	65	50	115

Tav. 2: Esplicite ed implicite: le cifre.

Si può affermare dunque che, tranne che in casi specifici, si usa con più frequenza l'esplicita[6]. Con *aconsejo*, di fronte a 19 casi di implicita, ci sono 45 occorrenze con la esplicita.

Alcuni esempi sono:

[5] Per questo rilievo, ringraziamo il professor Francesco Sabatini.

[6] Nello studio realizzato da George De Mello 1998, pp. 177-184, la frequenza, in termini assoluti, di subordinate esplicite rette da *aconsejar* è del 91% (10 occorrenze su 11); tale risultato è stato ottenuto dall'analisi del MC-NLCH (SAMPER PADILLA et alii 1998), coordinato da José Antonio Samper Padilla.

[8a] Te **aconsejo** que los publiques en el Rastro , tal vez por ahí le puedes sacar mejor precio
[8b] Mejor le **aconsejo** que no imite el ejemplo de la Doctora Cordero

Anche le altre forme verbali, tranne *aconseja* ed *aconsejan*, fanno parte di strutture con esplicita in numero più elevato. Alcuni esempi:

[9a] Este le **aconsejó** que la próxima vez que subiera al pulpito le pusiera un poco de vodka en el agua
[9b] Yo te **aconsejaría** que te informaras sobre el cirujano , el hospital y todo lo que tenga una relación

Un fattore che favorisce la scelta dell'esplicita è la presenza nella principale del clitico coreferente con il soggetto della subordinata.

[10a] Por fecha , ta bueno , pero , **te aconsejo** que esperes a septiembre debido a las lluvias , si es que este año las hay .
[10b] **Le aconsejé** que hiciera una modificación menor a la carta para que fuera legal .

Forma verbale	Esplicite			Implicite			Tot.
	nº occ.	con c. c.	senza c. c.	nº occ.	con c. c.	senza c. c.	
aconsejo	45	45		19	19		64
aconseja	01	1		12	2	10	13
aconsejamos	03	3		02		2	05
aconsejan	00			06	2	4	06
ha aconsejado	01	1		01	1		02
aconsejaba	01	1		01		1	02
aconsejé	02	2		01		1	03
aconsejó	04	4		03	3		07
aconsejaron	03	3		02	2		05
aconsejaría	05	5		02	2		07
aconsejaré	00			01		1	01
Totale	65	65		50	31	19	115

Tav. 3: Esplicite ed implicite: le cifre in rapporto alla presenza nella principale del clitico coreferente con il soggetto della subordinata.

Si osserva che in tutte le costruzioni con subordinata esplicita vi è il coreferente. Questo però non implica che, se c'è il coreferente, non si possa costruire la frase con la subordinata implicita. Entrambe queste osservazioni possono essere esemplificate con le occorrenze di *aconsejo*.

[11a] **Te aconsejo** que bajes los drivers para la ATI , porque no sé si la reconoce automáticamente .
[11b] **Les aconsejo** bajar el SpyBot , este programa , a diferencia del AD-AWARE , remueve " Efectivamente " la gama completa
[11c] Yo **te aconsejo** que vayas con algun mecanico para que le eche una miradita
[11d] Yo **te aconsejo** ir a un buen lugar de frenos , no se si tu conoces uno en especial ,

Come si nota agevolmente, in tutte le frasi si trova il clitico coreferente[7]. Va sottolineato che l'**assenza** nella principale del clitico pronominale coreferente è un fattore che determina la scelta della costruzione implicita.

[12a] El ritual original aconseja decorar un espacio con telas y cojines de colores suaves , hacer una especie de altar y encender una barrita

[12b] [...] las normas de cortesía **aconsejan** no rehusar un obsequio [...]

1.3.1 COREFERENTE NOMINALE. Come si è visto, nella quasi totalità dei casi la coreferenza al soggetto della subordinata viene effettuata da un pronome clitico che svolge la funzione di Oggetto Indiretto nella principale; ma conviene segnalare che tale pronome non è l'unico elemento che può essere coreferente. In realtà, la principale può anche avere un SN oggetto indiretto che funge anche da coreferente del soggetto della subordinata. Questa possibilità è molto meno frequente, come lo dimostra il fatto che nel corpus spagnolo si riscontrano soltanto tre casi di frase con SN coreferente. Tale irrilevanza quantitativa non permette di formulare nessuna ipotesi a proposito dell'uso della esplicita o della implicita. Comunque, per quanto riguarda le occorrenze trovate, le subordinate sono implicite:

[13] Resulta que mi cliente le había comprado una de sus bases de datos o servicio de spam . Yo **aconsejé** a mi cliente no hacerlo (por el efecto negativo que tiene sobre el usuario

1.3.2 SOGGETTO NON SPECIFICO. Nella frase complessa che ha come nucleo verbale *consigliare*, la proposizione principale presenta la possibilità di non avere nessun elemento coreferente del soggetto della subordinata. Tale possibilità apre due opzioni diverse. Nella prima, la subordinata ha un soggetto espresso; nella seconda, la subordinata non contiene nessun elemento che faccia riferimento ad un soggetto diverso dal morfema flessivo del verbo, che è in terza persona singolare.

La prima opzione si costruisce di necessità con subordinata esplicita, dato che il soggetto viene espresso appunto nella subordinata (si veda *supra*). La seconda opzione ci interessa in modo particolare, perché l'assenza di riferimenti al soggetto determina l'uso dell'implicita. Inoltre, si vedrà che con determinati verbi l'implicita è l'unica struttura possibile.

Una struttura con la principale senza coreferente insieme alla subordinata senza soggetto espresso presenta due possibilità:

(1) la principale ha soggetto, espresso generalmente, e la subordinata non ce l'ha.
(2) la principale non ha soggetto specifico e la subordinata neanche.

In entrambe le possibilità *aconsejar*, il verbo della principale, è in terza persona. Il fatto certo è che, in ambedue i casi, la subordinata non ha soggetto identificabile. Si tratta di un soggetto non specifico o generico.

Questi due casi trovano abbondante esemplificazione con la forma verbale *aconseja*, che in questo corpus regge subordinate implicite. In cinque casi il verbo è costruito in assenza di clitico; in quattro, oltre a non avere clitico coreferente, ha il *se* impersonale.

Nel corpus spagnolo si trovano occorrenze di tipo (1) sempre con la forma verbale al presente ed in terza persona, sia singolare (*aconseja*) che plurale (*aconsejan*). In tutte le occorrenze di questo tipo, la subordinata è implicita. Nella principale non ci sono clitici né altri elementi che facciano riferimento ad un eventuale soggetto. D'altra parte, dato che l'infinitivo è privo di morfemi di persona, è chiaro che il soggetto della subordinata non è specifico.

[7] Tranne un caso scritto in stile telegrafico.

(a) ***Aconseja***. Senza oggetto indiretto coreferente (né clitico né SN):

[14a] El asunto es que mi amigo **aconseja** cambiar si o si aceite , filtros , correa de distrubución.

[14b] El ritual original **aconseja** decorar un espacio con telas y cojines de colores suaves , hacer una especie de altar y encender una barrita

(b) ***Aconsejan***. Senza oggetto indiretto: né clitico coreferente né SN:

[15a] Los expertos **aconsejan** usar el mnemotécnico " ABC " en caso de accidentes : Ambulancia , Bomberos y luego Carabineros .

[15b] Artículo 17.- Si , tratándose de personalidades extranjeras o de visitas a países extranjeros , las normas de cortesía **aconsejan** no rehusar un obsequio , el Senador debe aceptarlo y ,

Per quanto riguarda il tipo (2), il verbo *aconsejar* è sempre in terza persona singolare e va preceduto dall'impersonale *se*. Nel corpus si trovano quattro occorrenze senza pronome clitico coreferente. Tutte hanno la subordinata implicita.

(c) ***Se aconseja***. Senza complemento indiretto coreferente (né clitico né SN):

[16a] servicio de Soporte Técnico de Panda Software , y en prevención de posibles encuentros con Blaster , **se aconseja** actualizar de inmediato las soluciones antivirus .

[16b] Me ha llegado un mensaje precioso donde **se aconseja** enfrentar a los problemas como sea , aun rompiendo el jarrón.

1.3.3 IL *SE* IMPERSONALE. La presenza del *se* (in italiano, del *si*) è uno dei meccanismi a disposizione della grammatica per indicare che la persona indicata dai morfemi flessivi verbali non è specifica.

Se basta per indicare che il soggetto della frase riceve un'interpretazione arbitraria, cioè è non specifico; ma non basta per indicare che è generico. Miguel Aparicio 1992, pp. 154-155, afferma che la "genericità" è collegata all'aspetto del verbo; di fatto, perché una frase riceva una lettura generica è necessario, oltre alla presenza di *se*, che il valore aspettuale sia imperfettivo.

Nel corpus, tutti gli esempi trovati hanno il verbo al presente con valore imperfettivo, e quindi possiamo ritenere che il "soggetto" è non specifico ed è generico. Miguel Aparicio sottolinea che *se* è un clitico privo di tratto di persona[8]; esso impedisce dunque la concordanza personale del verbo, che compare in 3ª persona perché questa è l'opzione non marcata.

1.3.4 IL *SE* IMPERSONALE ED IL VERBO *ACONSEJAR*. Con clitico coreferente nella principale, la subordinata può essere esplicita od implicita. L'unico esempio tratto dal corpus presenta la subordinata implicita:

[17] El repuesto para el vidrio no se encuentra en Chile por lo que **se me aconseja** volver al siguiente Lunes cuando se haya conseguido el vidrio .

Se la principale non ha il coreferente, di norma la subordinata è implicita. Con tale struttura, non sono specifici né il soggetto della principale — dato che c'è il *se* impersonale — né il soggetto della subordinata — visto che non c'è nessun elemento che vi faccia riferimento (nessun coreferente).

---

[8] Ciò nonostante, Cinque ritiene che *si* ha, come tutti gli elementi pronominali, un tratto di persona; tuttavia, è un tratto di persona non specificata, non referenziale: è un tratto incapace di selezionare da sé un referente specifico (Miguel Aparicio 1992, p. 161).

[18] nal de servicio de Soporte Técnico de Panda Software , y en prevención de posibles encuentros con Blaster , se **aconseja** actualizar de inmediato las soluciones antivirus .

1.3.5 SELEZIONE DELL'IMPLICITA CON SOGGETTO NON SPECIFICO. L'assenza di soggetto della subordinata e di qualsiasi riferimento ad esso, induce fortemente alla selezione dell'infinito nella subordinata [19a]. Se, comunque, la subordinata è esplicita, essa è costruita, in assenza di soggetto, con il *se* impersonale-passivo [19b]. Ora, se questo *se* è già nella principale [20a], il suo uso nella subordinata risulta, pur essendo grammaticale, molto forzato e cacofonico [20b]. Inoltre, quando il verbo della subordinata è riflessivo, l'uso del *se* impersonale nella subordinata viene impedito ed è agrammaticale [21].

[19a] Los expertos **aconsejan** usar el mnemotécnico " ABC " en caso de accidentes : Ambulancia , Bomberos y luego Carabineros .
[19b] Los expertos **aconsejan** que **se use** el mnemotécnico " ABC "...
[es. *a* trasformato in esplicita]

[20a] **Se aconseja** forrar molde exteriormente con alusa plas . (para que no penetre el agua del baño maría)
[20b] **Se aconseja** que **se forre** molde exteriormente con alusa plas
[es. *a* trasformato in esplicita]

[21] ***Se aconseja** que **se duche** antes de entrar en la piscina.

Il *se* della subordinata si rivela indubbiamente passivo quando l'oggetto dell'infinito è al plurale. Nell'operazione di trasformazione dall'implicita all'esplicita, se l'oggetto del verbo all'infinito è plurale, deve concordare in numero con il verbo dell'esplicita:

[22a] y en prevención de posibles encuentros con Blaster **, se aconseja** actualizar de inmediato las soluciones antivirus .
[22b] **.. , se aconseja** que **se actualicen** de inmediato las soluciones antivirus .
[come es. *a* trasformato in esplicita]

Anche se il corpus non ne offre esempi, abbiamo potuto riscontrare che i verbi riflessivi non ammettono l'esplicita con *se* [23b]. Anche gli intransitivi presentano frasi dubbiose [24b]. Inoltre alcune frasi con verbo transitivo e con oggetto non determinato non risultano accettabili.

[23a] Se aconseja lavarse las manos www.thyroid.com/sp/guide.html
[23b] *Se aconseja que se lave las manos
[es. *a* trasformato in esplicita][9]

[24a] Aparcamientos.- Se han habilitado los P-3, P-6 y P-8, todos en la Feria. Se aconseja ir a pie desde aquí. www.20minutos.es 24.06.2005
[24b] *Se aconseja que se vaya a pie desde aquí.
[es. *a* trasformato in esplicita]

[9] Le frasi con subordinata esplicita che hanno il verbo riflessivo sono agrammaticali:

[25a]	*Se aconseja ducharse.*	[25a']	**Se aconseja que se duche.*
[25b]	*Se aconseja ponerse corbata*	[25b']	**Se aconseja que se ponga corbata.*

Si noti che anche con verbo transitivo nella subordinata, se esso ha come oggetto un nome senza determinante, non si può costruire con esplicita:

[25c]	*Se aconseja llevar corbata*	[25c']	*??Se aconseja que se lleve corbata*
[25d]	*Se aconseja usar gafas de sol*	[25d']	**Se aconseja que se usen gafas de sol.*
[25e]	*Se aconseja comer fruta*	[25e']	*??Se aconseja que se coma fruta.*

A proposito dell'agrammaticalità (chiara nel caso dei verbi riflessivi) o dell'inaccettabilità della subordinata esplicita in queste condizioni sintattiche, importa mettere in rilievo il fatto che in spagnolo non sempre è ammessa (e meno ancora raccomandabile) la subordinata esplicita. Per poter affermare che la subordinata esplicita è più frequente (come è confermato in questo corpus) e che è raccomandabile, si deve porre come condizione la presenza del clitico coreferente. Al contrario, la subordinata implicita è più frequente, e quindi raccomandabile, in assenza, appunto, del coreferente, cioè quando il soggetto della subordinata non è specifico.

2. DISSIMMETRIE *CONSIGLIARE / ACONSEJAR*. Ne vanno contemplate almeno tre casi.

2.1 CON CLITICO COREFERENTE NELLA PRINCIPALE. In italiano, in questo caso, non è possibile la subordinata esplicita:

[26] ***Ti consiglio** che tu non esca con questa pioggia.

In spagnolo, se nella principale c'è il clitico coreferente del soggetto della subordinata, si usa di preferenza la subordinata esplicita[10]. Infatti, l'analisi del corpus generico NUNC spagnolo dei corpora.unito.it dimostra che l'esplicita viene usata molto più frequentemente con questa struttura (vedere tabella I e II di *aconsejar*).

[27] `Os aconsejo` `que utiliceis la anotacion por coordenadas , ya que si no podeis facilmente equivocaros .`

2.2 COSTRUZIONE PASSIVA DEL VERBO REGGENTE. Come è noto, la frequenza e le possibilità di uso della passiva perifrastica in italiano ed in spagnolo presentano differenze che coprono un raggio molto più ampio, e quindi non si circoscrivono soltanto al caso dei verbi *consigliare* ed *aconsejar*. Le particolari differenze tra questi due verbi rispetto alla costruzione passiva meritano una segnalazione perché presentano delle peculiarità interessanti.

I verbi come *consigliare* «permettono la costruzione del passivo impersonale» (*GGIC* II, p. 656). La subordinata, come si osserva nell'esempio, è infinitiva.

[28] È stato proposto/detto/deciso/ordinato/proibito di partire. *GGIC* II, p. 656

Nonostante ciò Renzi, quando riprende l'argomento della costruzione passiva (*GGIC* II, p. 672), specifica che nelle infinitive, nei casi in cui tale costruzione è ammessa, essa richiede la presenza del complemento indiretto del verbo. Negli esempi, Renzi indica come agrammaticale il verbo *consigliare* e come dubbioso il verbo *ordinare*.

[29] È stato *consigliato / ?ordinato di partire. *GGIC* II, p. 672

Le frasi sono invece grammaticali con la presenza dell'oggetto indiretto:

[30] A Gianni / Mi è stato consigliato / ordinato di partire. *GGIC* II, p. 672

Nel corpus abbiamo trovato, su 1000 *matches* di *consigliato*, 12 occorrenze di costruzioni passive con coreferente clitico, cfr. es. [31a], ed uno senza, es. [31b]:

[31a] `Non ho mai frequentato questo NG , però` **`mi`** `è stato` **`consigliato`** `di farlo perchè c'è da divertirsi`

[31b] `Spesso è stato` **`consigliato`** `in questo ng di alimentare le tarte nel modo più vario possibile`

---

[10] Questo fatto giustifica la "regola pratica" usata da diversi insegnanti di spagnolo, che consiste nel dire agli studenti italofoni che in spagnolo i verbi di influenza si costruiscono solo con la subordinata esplicita.

Nel corpus spagnolo non si è trovato nessun caso di passiva perifrastica del verbo *aconsejar* che regga una subordinata. In genere sono molto più frequenti le passive con il *se* di quelle perifrastiche. Di fatto, nel corpus si sono trovate cinque passive con *se*: quattro senza clitico coreferente ed una con clitico. In tutti i casi, la subordinata è infinitiva:

[32a] Se **aconseja** forrar molde exteriormente con alusa plas . (para que no penetre el agua del baño maría)

[32b] El repuesto para el vidrio no se encuentra en Chile por lo que se **me aconseja** volver al siguiente Lunes cuando se haya conseguido el vidrio .

2.3 FRASE COPULATIVA. La traduzione letterale in spagnolo dell'es. [33]

[33] **è consigliato** proteggere la terra con un foglio di plastica.

sarebbe agrammaticale: **{es/está} aconsejado proteger la tierra con ...* ; il corrispondente spagnolo della struttura *è consigliato* + infinito è il *se* impersonale: *se aconseja proteger la tierra con ....*

3. SIMMETRIE *CONSIGLIARE / ACONSEJAR*[11]: SOGGETTO NON SPECIFICO E GENERICO. In italiano, Renzi afferma che, se il soggetto della subordinata è impersonale, è molto più frequente la subordinata temporalizzata con l'impiego di *si* o della forma passiva (*GGIC* II, p. 644). L'esempio di Renzi è

[34] Il generale ordinò che si facesse saltare il ponte. *GGIC* II, p. 644

Il corpus NUNC Italiano I non offre nessun esempio di *consigliare* con *si* impersonale nella subordinata esplicita. Tutte le occorrenze con soggetto non specifico (senza coreferente) hanno la subordinata implicita. Quindi, quella che sembrava una differenza fra le due lingue si è rivelata invece un uso simmetrico. Alcuni esempi sono:

[35a] Il programmatore di HTML POP3 **consiglia** inoltre di attivare l' opzioni " mantieni copia sul server " ,

[35b] La seguente scaletta potrebbe subire modifiche , **si consiglia** quindi di consultare questa pagina per eventuali aggiornamenti .

In spagnolo, quando il soggetto è generico o non specifico, è più frequente l'implicita. Quando non è presente il clitico coreferente nella principale, si interpreta che il soggetto della subordinata è generico o non specifico. Di fatto, nei corpora, le occorrenze senza coreferente pronominale sono costruite con la subordinata implicita:

[36] Apreciado Nostromo , Te envío el estofado solicitado que según el autor , **aconseja** comer solo a mediodía !!!!

Le occorrenze con *se* impersonale hanno la subordinata implicita.

[37] Me ha llegado un mensaje precioso donde **se aconseja** enfrentar a los problemas como sea , aun rompiendo el jarrón.

4. CONCLUSIONI. Per quanto riguarda rispettivamente *consigliare* ed *aconsejar* le conclusioni sono pertanto le seguenti.

[11] Sono anche simmetriche le costruzioni di *consigliare* ed *aconsejar* come verbo dichiarativo, da un lato, e come verbo di influenza con soggetto espresso nella subordinata, dall'altro.

4.1 *CONSIGLIARE*. Due i problemi che si pongono:

(1) Si usa la subordinata esplicita retta da *consigliare* come verbo di influenza senza soggetto espresso nella subordinata?
• Nel corpus non si è trovato nessun caso.
• Le grammatiche segnalano però, in particolare, il possibile uso dell'esplicita con il *si* passivante nella subordinata.

(2) Si usa sempre l'introduttore *di* nelle implicite rette da *consigliare*?
• Sì, tranne che in frasi copulative con la struttura: *è* + *consigliato* + infinito, come è previsto da Renzi.
• Si noti però che, a differenza di quanto previsto da Renzi, i dati del nostro corpus forniscono al meno un esempio con copulativa e *di*.

4.2 *ACONSEJAR*. Tre i problemi che si pongono:

(1) Si usa la subordinata implicita retta dal verbo *aconsejar* come verbo di influenza senza soggetto espresso nella subordinata?
• Sì.

(2) È indifferente l'uso dell'esplicita o dell'implicita?
• No. Si deve specificare in quali contesti sintattici è preferito l'uso dell'implicita.
• I risultati ottenuti dall'osservazione delle occorrenze del corpus generico NUNC spagnolo sono chiari: (a) la presenza nella principale del clitico coreferente del soggetto della subordinata favorisce l'uso dell'esplicita – di fatto, nel corpus questo è il caso più frequente –; (b) l'assenza del clitico coreferente nella principale induce fortemente all'uso dell'implicita; (c) il *se* impersonale nella principale insieme all'assenza di clitici coreferenti determina l'uso dell'implicita.

(3) Quando è obbligatoria la struttura implicita?
• Quando si verifica la presenza nella principale del *se* impersonale, insieme all'assenza di clitici, ed inoltre il verbo della subordinata è riflessivo. Se si presentano queste condizioni, la frase con subordinata esplicita è agrammaticale.

BIBLIOGRAFIA.

BOSQUE - DEMONTE
1999 *Gramática descriptiva de la lengua española*, dirigida por Ignacio Bosque y Violeta Demonte, preámbulo de Fernando Lázaro Carreter, índices a cargo de Ma. Victoria Pavón Lucero, Madrid, Espasa-Calpe, 1999, 3 voll.

COSERIU
1988/92 Eugenio Coseriu, *Competencia lingüística. Elementos de la teoría del hablar*, elaborado y editado por Heinrich Weber, versión española de Francisco Meno Blanco, Madrid, Gredos, 1992 [Edizione originale: Eugenio Coseriu, *Sprachkompetenz: Grundzüge der Theorie des Sprechens*, Tübingen, Francke, 1988].

DELBECQUE - PAEPE
1988 *Estudios en honor del profesor Josse de Kock*, reunidos por N[icole] Delbecque y C[hristian] De Paepe, Leuven, Leuven University Press, 1998 "Symbolae Facultatis Littaerarum Lovaniensis. Series A" 25.

DISC → SABATINI - COLETTI 2003.

GGIC I → RENZI - SALVI et alii 1988; II → RENZI - SALVI et alii 1991; III → RENZI - SALVI et alii 1995.

MELLO
1988 George De Mello, *Verbos de influencia + cláusula / infinitivo con sujetos no correferenciales*, in DELBECQUE - PAEPE 1988, pp. 177-184.

MIGUEL APARICIO
1992 Elena De Miguel Aparicio, *El aspecto en la sintaxis del español: Perfectividad e impersonalidad*, Madrid, Ediciones de la Universidad Autónoma de Madrid, 1992.

RENZI - SALVI et alii
1988 *Grande grammatica italiana di consultazione*. Volume I, *La frase. I sintagmi nominale e preposizionale*, a cura di Lorenzo Renzi, Bologna, il Mulino, 1988.
1991 *Grande grammatica italiana di consultazione*. Volume II, *I sintagmi verbale, aggettivale, avverbiale. La subordinazione*, a cura di Lorenzo Renzi e Giampaolo Salvi, Bologna, il Mulino, 1991.
1995 *Grande grammatica italiana di consultazione*. Volume III, *Tipi di frase, deissi, formazione delle parole*, a cura di Lorenzo Renzi, Giampaolo Salvi e Anna Cardinaletti, Bologna, il Mulino, 1995.

SABATINI - COLETTI
2003 *Il Sabatini Coletti. Dizionario della lingua italiana*, diretto da Francesco Sabatini e Vittorio Coletti, Milano, Rizzoli Larousse, 2003, 1 vol. + 1 CD-ROM.

SAMPER PADILLA et alii
1998 *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*, preparado por José Antonio Samper Padilla, Clara Eugenia Hernández Cabrera y Magnolia Troya Déniz, Las Palmas de Gran Canaria, Universidad de Las Palmas de Gran Canaria, 1998.

TORRENTE SÁNCHEZ-GUISANDE
1998 Francisca Ángela Torrente Sánchez-Guisande, *Oraciones subordinadas sustantivas. Uso del indicativo, el subjuntivo y el infinitivo*, presentación de Carmen Martín Gaite, Firenze, Alinea, 1998 "Lingue d'Europa".

CORPORA DI RIFERIMENTO.

Corpora.unito.it http://www.corpora.unito.it/.

MC-NLCH http://listserv.rediris.es/cgi-bin/wa?A2=ind9901&L=infoling&P=670 (SAMPER PADILLA et alii 1998)

NUNC-ES Generic http://www.bmanuel.org/projects/ng-HOME.html.

NUNC-IT Generic I http://www.bmanuel.org/projects/ng-HOME.html.

PURA GUIL - MARGARITA BORREGUERO ZULOAGA
(Universidad Complutense de Madrid)

18. Comparative prototipiche in italiano e spagnolo. *I NUNC come base per l'analisi contrastiva.*

0. INTRODUZIONE. Le comparative prototipiche, come altre formule linguistiche cristallizzate, sono state oggetto di studio non soltanto della lessicografia, ma anche della sociologia, della teoria della letteratura, della teoria dell'argomentazione e della semantica (Amossy - Herschberg-Pierrot 2001). In questa ultima disciplina, soprattutto nelle recenti proposte della semantica cognitiva, le formule stereotipate sono considerate un indizio evidente del modo in cui ogni comunità linguistica percepisce e categorizza la realtà, per mezzo di generalizzazioni e semplificazioni che a volte sono eccessive.

Sono invece assai rari gli studi che si occupano di analizzare in modo contrastivo le caratteristiche formali e funzionali di questo tipo di strutture in diverse lingue, nonostante il consenso esistente riguardo alla loro importanza per l'analisi della caratterizzazione delle comunità socioculturali.

Questo lavoro costituisce un primo approccio contrastivo alle comparative prototipiche di base aggettivale dell'italiano e dello spagnolo, facendo tesoro della ricchezza di materiali che i NUNC mettono a nostra disposizione in entrambe le lingue.

1. QUALCHE OSSERVAZIONE SULLA STRUTTURA SINTATTICA. Dato che lo scopo di questo lavoro non è un approccio sintattico alle comparative prototipiche e che la loro struttura formale è stato l'aspetto che ha ricevuto una maggiore attenzione, in questo paragrafo ci limitiamo ad accennare alcune considerazioni che ci sono sembrate particolarmente rilevanti per l'analisi semantica e pragmatica che svolgeremo in seguito. Per ulteriori approfondimenti rimandiamo alla bibliografia.

1.1 VARIAZIONI FORMALI DELLE STRUTTURE COMPARATIVE. La struttura canonica delle comparative prototipiche di base aggettivale, cioè che modificano un aggettivo, è:

Aggettivo + *come* + SN

Alla stregua di Bosque 1999, p. 220 consideriamo che il complemento di paragone è in realtà un modificatore della testa aggettivale. I modificatori vengono tendenzialmente interpretati come quantificatori e si trovano in posizione preaggettivale (*molto alto, troppo vecchio*, *terribilmente suscettibile, eccezionalmente intelligente*), ma non esclusivamente: *pieno zeppo*, *povero in canna*, *stupido all'estremo, bugiardo matricolato, cretino patentato, golosa da morire, bello da impazzire,* ecc.

Dunque, anche se appaiono sempre in posizione post-aggettivale, i complementi comparativi degli aggettivi sono modificatori anche loro, dato che realizzano funzioni di quantificazione.

Senza entrare nel dibattito circa l'origine di queste strutture, dibattito che ruota intorno all'esistenza o meno di una predicazione ellittica formata da un verbo copulativo più l'aggettivo (Vietri 1990; Sáez del Álamo 1999), ci limiteremo a presentare le principali forme in cui possono apparire. Inizieremo la nostra analisi con una breve riflessione sulla diversa costituzione dei SN.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 309-322.

1.1.1 IL SINTAGMA NOMINALE TERMINE DI PARAGONE. I SSNN che costituiscono il termine di comparazione di queste strutture possono essere più o meno complessi. Ecco qui un elenco delle forme più frequenti: un nome proprio (cfr. es. [1a]), un SN senza articolo (es. [1b]), un SN con un articolo determinativo (es. [1c]) od indeterminativo (es. [1d]), un SN con uno o più modificatori aggettivali (es. [1e]) o forme participiali (es. [1f]), un SN con uno o più modificatori preposizionali (ess. [1g-h]), un SN modificato da una frase relativa (es. [1i]).

[1a] interessi squallidi di ipocriti falsi come Giuda[1]
[1b] Bologna, centro storico. Il nebbione denso come fumo
[1c] Una melodia antica come il mondo
[1d] Quelli che mi fanno essere contenta come una bambina
[1e] affidabile come un gatto randagio
[1f] dolce come una sfogliatella appena sfornata
[1g] goffo come un bambino con il pannolone
[1h] duro come un marciapiede di granito
[1i] ma anche Bonolis è simpatico come un gatto che si aggrappa alle palle nel tentativo di non cadere

1.1.2 LE POSSIBILI STRUTTURE COMPARATIVE. Alcune di queste espressioni ammettono le varianti sintattiche, correlative ai diversi gradi di comparazione espressi:

[2a] su vestido blanco como la nieve
[2b] su vestido tan blanco como la nieve
[2c] su vestido más blanco que la nieve

[3a] hai la zucca dura come un diamante
[3b] hai la zucca dura quanto un diamante
[3c] hai la zucca più dura di un diamante

[4a] La scusa vecchia come il mondo
[4b] trucchi vecchi quanto il mondo

In alcuni casi non è possibile usare l'una o l'altra di queste strutture indistintamente. Per esempio, tanto in spagnolo come in italiano

	possiamo dire	*ma non*
[5a]	eres más tonto que Abundio	*eres tonto como Abundio
		*eres tan tonto como Abundio
[5b]	es más infeliz que un cubo	*es infeliz como un cubo
[5c]	era libre como el viento	*era más libre que el viento
[6a]	È muto come un pesce	*è più muto di un pesce

Comunque, al margine di questi casi eccezionali, ciò che ci preme segnalare è che le variazioni della struttura non comportano nessuna modifica né dell'operazione di quantificazione né dal punto di vista semantico. Sia *hai la zucca più dura del diamante* sia *hai la zucca dura come un diamante* possono essere sostituite da *hai la zucca durissima*, senza che l'uso della comparativa di maggioranza implichi nessun incremento dell'intensificazione. Si tratta, di conseguenza, di un modo di "sfogare" le necessità espressive del parlante, una variazione cioè che ha la sua importanza sul piano pragmatico (sul quale torneremo più avanti), ma non sul piano semantico.

[1] Salvo diversamente avvisato, gli esempi in Courier in italiano sono tratti da NUNC-IT Generic, quelli spagnoli da NUNC-ES Generic; in Times, invece, sono gli *exempla ficta* e quelli tratti da altre fonti (segnalate).

Questa invariabilità semantica si produce anche in quei casi che presentano un significato figurato. In effetti, nel corpus costituito da Casadei 1996 nella sua ricerca semantica delle espressioni idiomatiche, troviamo il significato di 'ovvio, inconfutabile' attribuito ad ognuna delle seguenti espressioni:

[7a] essere chiaro come il sole — Casadei 1996, p. 425,
[7b] essere più chiaro del sole — Casadei 199, p. 428.

Vietri 1990, p. 154, segnala che le comparazioni prototipiche possono essere ridotte a strutture metaforiche più semplici, nelle quali scompare l'aggettivo od addirittura il *come,* e ne offre i seguenti esempi:

[8a] Max è docile come un agnello — Vietri 1990, p. 154,
[8b] Max è come un agnello — Vietri 1990, p. 154,
[8c] Max è un agnello — Vietri 1990, p. 154.

Anche noi abbiamo trovato diversi esempi di questo tipo nel NUNC,

[9] `Un governo di onesti è [raro] come un bordello di vergini`

ma non sempre queste riduzioni sono fattibili:

[10a] `uno è matto come un cavallo`
[10b] *uno è come un cavallo
[10c] *uno è un cavallo

La possibilità di riduzione può essere attribuita al grado di diffusione o di fissazione dell'immagine stereotipata (García Page 1996, p. 58).

1.1.3 IL VERBO. Il verbo italiano più usato in queste strutture è il verbo *essere*, mentre per lo spagnolo sono molto frequenti sia *ser* che *estar.* Quest'ultimo viene usato con quegli aggettivi che rinviano ad episodi o stati raggiunti dal soggetto ("predicati di stadio": cfr. Guil *i.s.*):

[11a] Está sordo como una tapia [ma non era così quando era giovane]
[11b] Está borracho como una cuba [ma oggi è sabato sera, lunedì non sarà più così].

Ciononostante, si tenga presente che ci sono altri aggettivi qualificativi che aspettualmente ammettono un doppio uso in spagnolo: sia come predicati individuali che possono essere descritti come stadi episodici (*María es rubia / María está rubia*), sia come predicati di stadio descritti come proprietà (*María está tranquila / María es tranquila*). E, naturalmente, ci possono essere anche oscillazioni diatopiche: nei NUNC troviamo frequentemente *es claro como el agua*, espressione usata in varietà latino-americane dello spagnolo, mentre nella varietà europea si adopera *está claro como el agua* (od *está más claro que el agua*).

Si noti inoltre che, a parte le motivazioni semantiche, qui vogliamo solo accennare che per avere l'eliminazione dell'avverbio *come* sembra necessario che il verbo sia *essere* o *ser,* e non *estar*. Invece la soppressione dell'aggettivo è possibile con i due verbi in entrambe le lingue:

[12a] `Es alto como un gigante`
→ *Es como un gigante* → *Es un gigante*
[12b] `Sei rapido come un fulmine`
→ *Sei come un fulmine* → *Sei un fulmine*

ma non

[13] Estás sordo como una tapia
→ *Estás como una tapia* → **Estás una tapia*

1.2 FUNZIONE DELLE COMPARATIVE PROTOTIPICHE: COMPARAZIONE O QUANTIFICAZIONE? Come prima accennato, secondo Ignacio Bosque, una comparativa prototipica è «una manifestación léxica estereotipada de la cuantificación de grado que varía en función del predicado intensificado» (Bosque 2004, p. cxxx). Quindi queste strutture sono, in primo luogo, quantificatori che indicano il grado massimo d'intensificazione di un predicato, sia questo un aggettivo od un verbo.

Questa funzione si vede chiaramente in

[14] La gente mangia carne e pensa che diventerà forte come un bue

dove non si comparano due entità diverse (*la gente* ed *un bue*), ma si stabilisce il grado di forza di una entità (*la gente*) per analogia con la forza di un'altra entità (*un bue*). In questo caso, *il bue* si considera un'entità che rappresenta il grado massimo di forza od almeno un grado molto elevato.

La differenza è ancora più ovvia se si comparano queste due occorrenze:

[15a] Nel prima di far l' amore con una fanciulla stavi mezz' ora a guardarle la pelle e la lucentezza degli occhi per capire se era sana come un pesce o infettiva come una bomba battereologica ambulante

[15b] La donna è come il pesce : tolta la testa , il resto tutto buono

Nel primo caso si tratta di una comparazione prototipica in cui non si paragona *una fanciulla* ed *un pesce*, ma si stabilisce intensificativamente il grado di salute della fanciulla; invece, il secondo è un caso di comparazione propria, dove si mettono a confronto, anche se parodicamente, due entità.

Inoltre, a differenza delle comparative proprie, in cui è sempre possibile permutare le posizioni e cambiare i modificatori, senza un mutamento di significato, come in

[16] Pietro è più furbo di Paolo → Paolo è meno furbo di Pietro,

in alcune comparative prototipiche queste permutazioni non sono possibili:

[17a] Questo film è lungo come la fame
→ **La fame è meno lunga di questo film*

[17b] Ana es más lista que el hambre
→ **El hambre es menos lista que Ana*

Le comparative di inferiorità risultanti in entrambi i casi sono sequenze grammaticali, ma semanticamente e pragmaticamente inadeguate.

La funzione intensificatrice è particolarmente evidente nella seguente occorrenza, in cui il parlante sviluppa il paragone in seguito con abbondanza di particolari:

[18] È uscita una vecchietta ... brutta .. molto brutta .. bassa .. magra come un chiodo , sembrava che non avesse carne , ma solo ossa ricoperte di pelle

Una prova ulteriore della funzione di quantificazione e d'intensificazione di questa struttura è la sua incompatibilità con gli avverbi di grado o con il superlativo, in modo da evitare la ridondanza che comporterebbe una doppia quantificazione:

[19a] es astuto como zorro viejo
[19b] *es muy astuto como zorro viejo
[19c] *es astutísimo como zorro viejo

[20a] è lento come una tartaruga

[20b] *è molto lento come una tartaruga

In italiano, però, abbiamo trovato espressioni come

[21a] `sono soddisfattissimo come un riccio`
[21b] `efficientissima come un' infirmiera svizzera`
[21c] `è stabilissima come una roccia`

che magari possono essere una spia di una desemantizzazione in corso del suffisso *-issimo* in questa lingua.

Si osservi pure che il suo valore intensificatore permette a questo costrutto di stabilire un rapporto paradigmatico con altre espressioni superlative:

[22a] `è duro come un macigno`
[22b] è durissimo
[22b] è molto duro

Solo in quei casi in cui si è sviluppato un significato idiomatico, la sostituzione non è più possibile:

[23a] `essere asciutto come l'esca`

non può essere sostituito da

[23b] *essere asciuttissimo

visto che questa espressione significa 'non avere denaro' (Casadei 1996, p. 425).

In sintesi, la peculiarità della comparativa prototipica risiede nel fatto che il termine di paragone svolge il ruolo di "misuratore" della proprietà espressa dall'aggettivo, rappresentandone il grado massimo. Si tratta in tutti i casi di un'immagine iperbolica. Da quest'equivalenza si ottiene un effetto d'intensificazione, ed in questo modo il proposito intensificatore prevale su quello puramente comparativo (Sáez del Álamo 1999, p. 1162).

2. CARATTERISTICHE SEMANTICHE DELLE COMPARATIVE PROTOTIPICHE. Saranno qui presi in considerazione gli aggettivi, i determinanti, e le entità prototipiche e termini di paragone.

2.1 GLI AGGETTIVI. Gli aggettivi che occorrono in queste strutture devono ammettere la quantificazione di grado, quindi devono denotare una proprietà graduabile:

[24a] `alto come una montagna`
[24b] `una scena vecchia come il mondo`
[24c] `fuerte como un toro`

Ciò nonostante, gli aggettivi di colore, che sono tipicamente qualificativi non graduabili, compaiono frequentemente in queste strutture:

[25a] `caffè nero come una notte senza luna`
[25b] `più bianca di un fantasma fiabesco`
[25c] `Dicono che nel folto de le` [sic] `chiome voi abbiate una ciocca rossa come una fiamma`

Allo stesso modo, gli aggettivi di relazione non sono graduabili, ma in alcuni casi possono venir ricategorizzati come qualificativi e far parte di una comparazione prototipica:

[26] `è più papista del papa`

È pure interessante sottolineare che questi aggettivi graduabili, che sono per natura relativi, sono usati in modo assoluto in queste costruzioni, visto che in esse si esprime non la "norma" ma il grado massimo della proprietà.

Si tratta in tutti i casi di aggettivi con un alto indice di frequenza nell'uso della lingua parlata. Fanno riferimento alle aree concettuali che esigono maggiore espressività: aspetto fisico, capacità intellettuali, età, attributi morali, ecc. (Ortega Ojeda 1990, p. 734).

Non è infrequente trovare la stessa comparazione con aggettivi che possono considerarsi sinonimi o varianti lessicali più o meno sinonimiche:

[27a] magro / secco come un chiodo
[27b] contento / felice come un bambino
[27c] cieco / miope come una talpa

Inoltre bisogna segnalare che questi aggettivi non si usano sempre nel loro significato letterale, ma è possibile che mettano in gioco contemporaneamente un significato figurato o metaforico rispetto all'entità di cui si predica la proprietà:

[28] se avesti letto il messaggio senza esserti chiuso come un riccio nelle tue convinzioni

In questo caso, la persona in questione è considerata chiusa nel senso metaforico di 'poco disponibile ad ascoltare le opinioni e le idee degli altri', ma la proprietà ha una denotazione fisica inerente al termine di paragone, *un riccio*: per difesa si avvolge completamente a palla.

[29] Juan es más agarrado que un chotis

In questo caso, *agarrado* è usato letteralmente per far riferimento al *chotis* (ballo di coppia del quale si dice che si deve ballare sopra un mattone), ma è usato figuratamente nel senso di 'tirchio', 'avaro' in riferimento a Juan.

2.2 I DETERMINANTI. I SSNN trovati nel corpus nella funzione di termine di paragone offrono sistematicamente una lettura ricollegabile in qualche modo alla categorialità, a seconda dei determinanti adoperati.

(a) Nei sintagmi privi di determinante si designa la categoria in modo astratto, in quanto concetto:

[30a] suave como terciopelo
[30b] denso come fumo

(b) I SSNN indeterminativi al singolare, con o senza specificazione attributiva, che costituiscono il gruppo più numeroso, in questi costrutti possiamo considerarli generici (d'accordo con la proposta di lettura fatta da Korzen 1996, p. 389):

[31a] affilata come una lama
[31b] borracho como una cuba
[31c] acido come una limonata senza zucchero

(c) Nei SSNN determinativi al plurale si fa la lettura generica, con rinvio alla categoria vista come classe aperta di entità numerabili:

[32a] fredda come i pesci
[32b] viejo como los dinosaurios

(d) I SSNN determinativi al singolare che abbiamo trovato rinviano ad "entità uniche" di notorietà generale,

[33a] antica come il mondo
[33b] caliente como el inferno

oppure alla categoria, nel caso di nomi non numerabili astratti o concreti o di nomi numerabili,

[34a] lungo come la fame
[34b] liscio come l'olio
[34c] blanco como la nieve
[34d] vecchio come il cucco
[34e] nero come la notte

oppure ancora alla sottocategoria, nel caso di nomi numerabili con specificazioni attributive,

[35] liscio come il culetto di un bebè

È vero che abbiamo trovato anche occorrenze di SSNN determinativi al singolare esprimenti una individuazione,

[36a] alto come la torre di Pisa
[36b] más pesado que el cuñado de Rocky

ma si tratta sempre di entità assunte dal parlante come il paradigma superlativo della proprietà predicata, e presentate quindi come il suo prototipo, analogamente a quello che succede quando viene adoperato un nome proprio, che rimanda ad un individuo ma solo in quanto convenzionalmente considerato il rappresentante della proprietà in questione:

[37a] interessi squallidi di ipocriti falsi come Giuda
[37b] más negro que Pelé

2.3 ENTITÀ PROTOTIPICHE E TERMINI DI PARAGONE. È chiaro che, sul piano cognitivo, le comparazioni costituiscono un efficace espediente per capire – e far capire – ciò che non è noto tramite ciò che è noto. Nel caso delle comparazioni prototipiche, si predica di un'entità una proprietà tramite il paragone con un'altra entità che si considera il rappresentante migliore della proprietà in questione.

In altre parole, all'interno di una comunità linguistica ed in modo più o meno rigido e convenzionale, all'entità assunta a termine di paragone viene attribuita in grado massimo la proprietà designata.

[38a] astuto come una volpe
[38b] Un giovane carabiniere di leva con la faccia bianca come un cencio
[38c] fa che sia breve come un fiocco di neve
[38d] ecco che la stampante parte contenta come una pasqua

Possiamo scomporre il processo eseguito dal parlante nelle seguenti fasi: **(1)** l'intenzione è quella di predicare in modo superlativo una proprietà dell'entità A; **(2)** si seleziona una entità B, in un mondo possibile, nel cui stereotipo (inteso come l'insieme di tratti o proprietà caratteristiche di un'entità: cfr. Hurford - Heasley 1983) appare la suddetta proprietà; **(3)** si presenta questa entità B come prototipo della proprietà, cioè, si considera che tale proprietà è centrale ed appare nella sua massima gradazione nell'entità B; **(4)** si stabilisce l'equivalenza tra le entità A e B, ottenendo come risultato l'intensificazione della proprietà attribuita ad A.

2.3.1 CARATTERISTICHE DELL'ENTITÀ B, TERMINE DI PARAGONE. L'entità deve far parte delle conoscenze condivise dal parlante e dal suo interlocutore per almeno due ragioni:

(1) perché fissata all'interno di un'espressione tradizionale (un cliché) del tipo *más feo que Picio, más viejo que Matusalén, limpio como los chorros del oro, sano come un pesce, paziente come Giobbe*, nelle quali buona parte dei parlanti non riconoscono più la motivazione semantica, proveniente in molti casi da allusioni a personaggi biblici e storici, aspetti della vita contadina, antichi costumi, ecc., ma ciò nonostante continuano ad adoperarle perché fanno parte della loro tradizione linguistica;

(2) perché fa riferimento ad aspetti dell'esperienza quotidiana comune: aspetto fisico di certe entità (*rosso come un peperone, rojo como un tomate*), valutazione di alcuni fenomeni meteorologici (*forte come un tuono, veloce come un fulmine, bello come il sole, chiaro come il giorno*), giudizi su fatti culturali o sociali (*semplice come la pappa al pomodoro, más contento que niño con zapatos nuevos*).

Comunque i limiti tra questi due tipi di conoscenze non sono chiari. Molte di quelle acquisite dall'esperienza diretta con l'ambiente fisico possono perdere la loro motivazione per i parlanti che non abitano più in quell'ambiente. Questo è accaduto con le conoscenze che riguardano il comportamento degli animali. Per esempio, molti parlanti che hanno sempre vissuto in città non possono più spiegarsi perché si dice *testardo come un mulo, furbo come una volpe, cobarde como una gallina, paziente come un cavallo*, ecc. In questo modo, le comparazioni diventano cliché convenzionali che i parlanti usano senza capirne veramente il significato.

D'altronde, l'arbitrarietà che regge la scelta dell'elemento assunto come termine di paragone in queste comparazioni tradizionali si palesa contrastivamente: perché il rappresentante migliore del colore rosso è il peperone per gli italofoni ed il pomodoro per gli ispanofoni? Inoltre:

[39a] Sordo come una campana
[39b] Sordo como una tapia

[40a] Sano come un pesce
[40b] Sano como una manzana

Si tratta ovviamente di scelte fossilizzate. Comunque, indipendentemente dal fatto che l'utente conosca o meno l'entità termine di paragone, sarà in grado di estrarre il contenuto encomiastico che la struttura trasmette.

Le nostre ricerche nei NUNC hanno offerto la possibilità di confermare un'ipotesi: la vitalità di queste costruzioni si manifesta specie nella creazione di nuovi termini di paragone. Questa creatività si osserva tanto negli sviluppi enfatici di espressioni stereotipate convenzionali,

[41a] `sei cieco come una talpa a mezzanotte`
[41b] `tutto è sempre stato così chiaro come una mattina d' agosto alle prime luci del sole`

quanto nella selezione di un'entità nuova che non costituisce il rappresentante convenzionale della proprietà nella comunità linguistica,

[42a] `A parte il fatto che cerchietto non vuole portarmi a mare e mi sento bianca come una mozzarella`
[42b] `atroce come una guerra di assiri nell' antica babilonia` [sic]

Dunque possiamo aggiungere due modalità ulteriori in cui un'entità è presentata come parte delle conoscenze condivise:

(3) Il parlante sceglie creativamente un'entità od una situazione e la presenta come prototipo della proprietà che vuole predicare.

Queste entità fanno parte delle conoscenze culturali di almeno una parte della comunità linguistica nella quale normalmente non vengono presentate come paradigma di queste proprietà:

[43a] lungo come un discorso di Cossiga
[43b] más contento que Geppeto [sic] con una Black&Decker
[43c] más negro que el sobaco de un escarabajo
[43d] más pesado que el cuñado de Rocky
[43e] el conductor es más negro que Pelé

In queste espressioni il parlante presuppone che l'interlocutore è capace di identificare determinate entità del mondo (reale o fittizio) come Cossiga, Geppetto, un trapano Black&Decker, Rocky ed il suo cognato, e che ha una qualche conoscenza circa le qualità e gli atteggiamenti più tipici di queste entità.

Ovviamente molte di queste conoscenze si circoscrivono ad un àmbito spaziale e temporale estremamente ristretto e ciò spiega il carattere effimero di queste espressioni: fra qualche anno nessuno si ricorderà più di Rocky ed i Black&Decker saranno stati sostituiti da altri strumenti più sofisticati. Speriamo però che almeno Geppetto non faccia la stessa fine.

(4) La creatività del parlante raggiunge il punto massimo quando l'entità o la situazione non esiste in nessun mondo, né fittizio né reale, ma viene creata appositamente per la costruzione comparativa:

[44a] más pesado que corbata de plomo
[44b] más pesado que tanque a pedales
[44c] más feliz que perro con dos colas

2.3.2 CENTRALITÀ DELLA PROPRIETÀ RIGUARDO ALL'ENTITÀ. Parlando delle possibili variazioni formali di questa struttura abbiamo fatto allusione al fenomeno della riduzione che, mediante la soppressione dell'aggettivo e del *come*, dà luogo a strutture metaforiche del tipo *Max è un agnello*, *Eres un ángel*.

Diversi studiosi hanno accennato al grado di diffusione o fissazione dell'immagine come causa di questa diversità di comportamento. Vogliamo aggiungere che, a nostro avviso, anche la maggiore o minore centralità della proprietà gioca un ruolo importante.

In effetti, riguardo alle entità che vengono scelte come rappresentanti prototipici della proprietà sono possibili tre casi:

(a) si tratta di una proprietà centrale o tipica di questa entità, perciò facilmente identificabile.

Vediamo il seguente esempio:

[45] Devo essere rossa come un peperone!

L'immagine prototipica che gli italiani hanno del peperone è formata, tra gli altri tratti, dal colore rosso. Si tratta di una proprietà centrale. Ciò spiega le seguenti possibili riduzioni:

[46a] Devo essere come un peperone!
[46a] Devo essere un peperone!

(b) si tratta di una proprietà dell'entità ma non centrale.

Vediamo il seguente esempio:

[47] matto come un cavallo

Certamente, i cavalli possono reagire in modi poco comprensibili per gli umani, ma questo non significa che la pazzia sia il tratto che meglio definisce i cavalli. Questo ci impedisce di dire

[48a] *è come un cavallo
[48b] *è un cavallo

per indicare al nostro interlocutore che la persona di cui parliamo è pazza.

Ciò nonostante, l'interprete identifica la struttura di intensificazione (aggettivo + *come* + SN) ed è capace di riconoscere l'iperbole, il grado massimo attribuito alla proprietà, anche se non direbbe mai che quella proprietà si addice a quella entità.

(c) si tratta di una proprietà che non soltanto non fa parte dell'entità, ma può essere un prototipo della qualità opposta; in questo caso abbiamo come risultato un uso antifrastico con effetto parodico: la qualità rappresentata dal termine di paragone è antitetica riguardo alla qualità espressa dall'aggettivo (Ortega Ojeda 1990).

Vediamo i seguenti esempi:

[49a] sei sveglio come una cozza bollita
[49b] simpatico come una zecca sul culo
[49c] affidabili come una lotteria
[49d] espressivo come un macigno

Tra il primo ed il secondo termine di paragone ci deve essere una somiglianza fisica od attitudinale. Inoltre, per l'uso degli aggettivi dimensionali, è necessaria una proporzione di dimensione e formato. Si può dire di una persona che è *alta come una giraffa*, ma più difficilmente si dirà che è *alta come una montagna*. A meno che si tratti di un'entità non fisica, che ammette più facilmente qualsiasi termine di comparazione:

[50] Ovviamente è una bufala grande come una casa

D'altra parte è necessario che esista comunque una sproporzione che permetta di riconoscere che siamo davanti ad un'iperbole,

[51] il tono della voce gelido come un iceberg

in caso contrario la costruzione risulta ambigua:

[52] me ha salido un grano como un garbanzo

3. DIMENSIONE PRAGMATICA. L'intenzione comunicativa più evidente nell'uso di queste strutture è la volontà di esaltare l'attribuzione di una determinata proprietà, mediante un'iperbole. Ma accanto a questa intenzione di base è facile scoprire la volontà di compiere questa esaltazione aggiungendo una nota umoristica, ingegnosa, a volte ironica. In questo modo il parlante incrementa la propria faccia positiva, presentandosi al suo interlocutore come una persona brillante, creativa, capace di svelare il lato umoristico della realtà; ma al tempo stesso, valorizza la faccia positiva del destinatario, visto che il parlante presuppone che egli abbia delle competenze necessarie per la decodifica di queste espressioni; quindi, anche lui è dotato di senso dell'umorismo e della rapidità d'ingegno necessari per capirle.

Questa finalità ludica si osserva in modo particolare nei casi in cui le comparative, convenzionali o creative, sono allungate con nuovi elementi che non apportano niente all'iperbole e che costituiscono segni evidenti di quella necessità di sfogare la propria espressività di cui si parlava prima. Alcuni esempi dai corpora sono:

[53a] più fredda di una pizza albanese
[53b] più pesante di una mucca armena
[53c] più bianca di un fantasma fiabesco
[53d] più numerosi di una folla di cinesi ad un matrimonio
[53e] più acida di una zitella scaduta e irrancidita
[53f] más pesado que Pavarotti vestido de buzo
[53g] más tonto que un mosquito lobotomizado

Per questo motivo il linguaggio dei giovani ed i registri informali in genere sono particolarmente adatti allo studio delle comparazioni prototipiche ed in questo senso il NUNC costituisce un corpus privilegiato vista la quantità delle interazioni tra giovani (al limite della chat: cfr. qui Corino ¶ 13) tra i testi che raccoglie. In questo contesto comunicativo il carattere informale e poco curato dei dialoghi giovanili è rafforzato dalla velocità e dall'immediatezza imposte dal mezzo tecnico (a differenza della conversazione faccia a faccia, nelle chat non c'è il tempo non solo per pianificare gli interventi, ma neanche per correggere o modificare i propri enunciati, una volta emessi, nello stesso turno di parola).

Uno studio di questo tipo eseguito con un altro tipo di materiale, per esempio dizionari o gli elenchi raccolti negli studi specializzati, ci avrebbe offerto un panorama ben diverso della realtà d'uso, della vitalità della costruzione e delle espressioni effettivamente usate.

4. PROSPETTIVE INNOVATIVE PER L'ANALISI CONTRASTIVA. Lo scopo della nostra ricerca, della quale qui abbiamo presentato soltanto il punto di partenza, è ovviamente un'analisi contrastiva della forma, funzione ed uso di queste strutture nelle lingue spagnola ed italiana. Comunque i primi risultati ottenuti sono fortemente condizionati dal corpus utilizzato come base per le nostre ricerche. Per quanto riguarda lo studio della lingua italiana il NUNC si è rivelato uno strumento estremamente utile per la consultazione, ben etichettato ed abbastanza ampio, per la parte spagnola invece non è ancora finito il processo di schedatura, imprescindibile per agevolare la ricerca linguistica, nel quale inoltre bisognerebbe tenere conto delle varietà di spagnolo della penisola iberica.

In base agli esempi estratti fino ad ora, possiamo dire che si può apprezzare un'importante differenza nel tipo di costruzioni comparative usate dai giovani italiani e dai giovani ispanofoni. I primi tendono a usare più frequentemente le comparazioni più convenzionali del tipo *sano come un pesce, pieno come un uovo, felice come un bambino,* siano queste cliché di cui non si conoscono più i motivi della comparazione, siano ancora vive come comparazioni effettive nella coscienza del parlante. Non si trovano invece riferimenti a personaggi biblici (solo due casi di Giuda) o storici che sarebbero invece frequenti in altre fasce di età. Esiste, accanto a queste, un buon numero di comparazioni originali, create a partire da riferimenti culturali.

Nel corpus spagnolo invece non abbiamo ancora trovato espressioni convenzionali del primo e secondo tipo menzionati sopra, ma compare un significativo numero di strutture in cui si fa riferimento ad una situazione o personaggio della realtà culturale odierna o ad una situazione nella quale è presente uno di questi personaggi ma estrapolato in un contesto estraneo, producendosi così la voluta comicità.

[54a] eres más lento que dejar a la Barbie embarazada
[54b] más triste que Adán en el día de las Madres
[54c=43b] más contento que Geppeto [sic] con una Black&Decker
[54d] mas lento que la vuelta ciclista a España en Cyclostatic
[54e] Eres más pesado que una reposición de los mejores momentos de la carta de ajuste

Inoltre, i giovani ispanofoni sembrano propensi a inventare termini di paragone completamente assurdi ed inesistenti, come la *corbata de plomo*, il *tanque a pedales*, il *perro con dos colas*. Ovviamente ciò manifesta un desiderio di originalità, di affermazione della propria personalità tramite il discorso, di mostrarsi vivaci, acuti, ingegnosi nell'uso del linguaggio. Le comparative prototipiche convenzionali non possono far parte di questo gioco perché creano l'effetto contrario troppo conformistico e sono appunto quasi inesistenti in questo registro linguistico.

Ripetiamo, ciò nonostante, che queste impressioni hanno bisogno di un'ulteriore conferma ed aspettiamo impazienti l'ampliamento del corpus spagnolo del NUNC, convinte della sua utilità come strumento di ricerca linguistica.

BIBLIOGRAFIA.

ÁLVAREZ DE MIRANDA - POLO

2002 *Lengua y diccionarios. Estudios ofrecidos a Manuel Seco*, reunidos por Pedro Álvarez de Miranda y José Polo, Madrid, Arco Libros, 2002.

AMOSSY - HERSCHBERG-PIERROT

1997 Ruth Amossy - Anne Herschberg-Pierrot, *Stéréotypes et clichés*, Paris, Nathan, 1997 [Trad. sp. di Lelia Gándara, *Estereotipos y clichés*, Buenos Aires, Eudeba, 2001].

BENINCÀ et alii

1996 *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*, a cura di Paola Benincà, Guglielmo Cinque, Tullio De Mauro e Nigel Vincent, Roma, Bulzoni, 1996 "Università di Roma La Sapienza, Dipartimento di scienze del linguaggio".

BOLSHAKOV - GALICIA HARO

2002 Igor A. Bolshakov - Sofía N. Galicia Haro, *Frasemas con* como *en español*, online a `http://terral.lsi.uned.es/ia-mlia/iberamia2002/papers/mlia02.pdf`

BOSQUE

1999 Ignacio Bosque, *El sintagma adjetival. Modificadores y complementos del adjetivo. Adjetivo y participio*, in BOSQUE - DEMONTE 1999, vol. I. pp. 217-230.

2004 Ignacio Bosque, *Presentación*, in BOSQUE et alii, pp. xvii-xxviii.

BOSQUE - DEMONTE

1999 *Gramática descriptiva de la lengua española*, dirigida por Ignacio Bosque y Violeta Demonte, preámbulo de Fernando Lázaro Carreter, índices a cargo de Ma. Victoria Pavón Lucero, Madrid, Espasa-Calpe, 1999, 3 voll.

BOSQUE et alii

2004 *Redes: Diccionario combinatorio del español contemporáneo*, a cura di Ignacio Bosque, Madrid, SM, 2004.

CASADEI

1995 Federica Casadei, *Per una definizione di 'espressione idiomatica' e una tipologia dell'idiomatico in italiano*, in "Lingua e Stile" II (1995) 335-358.

1996 Federica Casadei, *Metafore ed espressioni idiomatiche: uno studio semantico sull'italiano*. Roma, Bulzoni, 1996 "Università di Roma La Sapienza, Dipartimento di scienze del linguaggio".

CORINO

¶ 13 Elisa Corino, *NUNC est disputandum. Aspetti della testualità e questioni metodologiche*, in questo volume, pp. 225-252.

DE GIOIA
1994 Michele De Gioia, *Chiaro come il sole a mezzogiorno: una classe di avverbi idiomatici dell'italiano*, in "The Linguist" XXXIII (1994) 220-225.
1994a Michele De Gioia, *Sur quelques comparaisons d'adverbes figés de l'italien et du français*, in "Lingvisticae investigationes" XVIII (1994)[1] 89-119.

DE GIOIA - MARQUES-RANCHHOD → MARQUES-RANCHHOD - DE GIOIA.

DE MAURO - VOGHERA
1996 Tullio De Mauro - Miriam Voghera, *Scala mobile. Un punto di vista sui lessemi complessi*, in BENINCÀ et alii 1996, pp. 99-131.

GARCÍA-PAGE,
1996 Mario García-Page, *Más sobre la comparación fraseológica en español*, "Lingüística española actual" XVIII (1996)[1] 49-77.

GUIL
i.s. Pura Guil, *Sull'aspetto degli aggettivi*, intervento al VI Convegno Internazionale SILFI *Tradizione & Innovazione: la linguistica e filologia italiana alle soglie di un nuovo millenio*, Gerhard-Mercator-Universität Duisburg, 28 giugno - 2 luglio 2000, in corso di pubblicazione negli *Atti*.

HURFORD - HEASLEY
1983 James R[aymond] Hurford - Brendan Heasley, *Semantics: A Coursebook*, Cambridge, Cambridge University Press, 1983. [Trad. sp. di Elena de Miguel e Isabel López Fraguas, *Curso de Semántica*, Madrid, Visor, 1988].

KLEIBER
1990 Georges Kleiber, *La sémantique du protoype: categories et sens lexical*, Paris, PUF, 1990 [Trad. sp. di Antonio Rodríguez Rodríguez, *La semántica de los prototipos. Categoría y sentido léxico*, Madrid, Visor, 1995].

KORZEN
1996 Iørn Korzen, *L'articolo italiano tra concetto ed entità. Uno studio semantico-sintattico sugli articoli e sui sintagmi nominali italiani con e senza determinante – con un'indagine particolare sulla distribuzione del cosiddetto "articolo partitivo"*. Vol. I. *Considerazioni preliminari. Il sintagma nominale senza determinante*, Vol. II. *I sintagmi nominali con articolo. Conclusioni*, København, Museum Tusculanum Press, 1996 "Études Romanes" 36.

LÓPEZ GARCÍA
1994 Ángel López García, *Gramática del español.* I. *La oración compuesta*, Madrid, Arco Libros, 1994.
1994a Ángel López García, *Las expresiones comparativas*, in López García 1994, pp. 209-253.

MARQUES-RANCHHOD - DE GIOIA
1996 Elisabete Marques-Ranchhod - Michele De Gioia, *Comparative Romance Syntax. Frozen Adverbs in Italian and in Portuguese*, in "Lingvisticae investigationes" XX (1996)[1] 33-85.

MILLÁN
2002 José Antonio Millán, *"El mundo entero le saldrà al encuentro". Las comparaciones en sus repertorios*, in ÁLVAREZ DE MIRANDA - POLO 2002, pp. 183-197.

ORTEGA OJEDA
1990 Gonzalo Ortega Ojeda, *Comparaciones estereotipadas y superlatividad*, in *Actas del Congreso Sociedad Española de Lingüística. XX Aniversario*, a cura di Mari Ángeles Álvarez, Madrid, Gredos, 1990, vol. II., pp. 729-737.

SÁEZ DEL ÁLAMO
1999 Luis Sáez del Álamo, *Los cuantificadores: las construcciones comparativas y superlativas*, in BOSQUE - DEMONTE 1999, vol. I pp. 1129-1188.

VIETRI
1990 Simonetta Vietri, *On Some Comparative Frozen Sentences in Italian*, in “Lingvisticae investigationes” XIV (1990)[1] 149-174.

CORPORA DI RIFERIMENTO.

NUNC-ES Generic http://www.bmanuel.org/projects/ng-HOME.html.

NUNC-IT Generic http://www.bmanuel.org/projects/ng-HOME.html.

MILENA BINI - ALMUDENA PERNAS - PALOMA PERNAS
(Universidad Complutense Madrid)

19. Apprendimento / insegnamento delle collocazioni dell'italiano. *Con i NUNC è più facile.*

0. INTRODUZIONE. L'obiettivo del nostro lavoro è stato esplorare le potenzialità dei NUNC nel campo dell'insegnamento e dell'apprendimento dell'italiano come lingua straniera (LS/L2). I corpora elettronici sono infatti i luoghi deputati a fornire i materiali ideali per l'insegnamento / apprendimento di una seconda lingua intesa come mezzo di comunicazione, dato che si tratta di collezioni di testi parlati e/o scritti che rispecchiano l'uso reale della lingua in contesti concreti e variati. Ed i NUNC, tanto quello generale come i NUNC specialistici (cucina, motori, fotografia), rientrano in questa categoria. La nostra ricerca si è incentrata sulle collocazioni, su come reperirle nei NUNC e come aiutare gli studenti ad apprenderle.

1. LE COLLOCAZIONI. Le collocazioni, come si sa, sono «sequenze di parole che tendono a presentarsi in combinazioni stabili tra loro e privilegiate» (Simone 1990, p. 440), ma diversamente da quanto avviene nelle frasi idiomatiche le parole che le compongono non perdono il loro significato autonomo nè la loro funzione sintattica, per questo passano inosservate. Sono combinazioni frequentissime che si sono fossilizzate per esprimere un determinato significato complesso ma apparentemente non c'è niente che spieghi la loro formazione, per questo le collocazioni hanno un chiaro carattere idiosincratico. Ci sono varie categorie di collocazioni[1]: nome + verbo (*la macchina sbanda, la tempesta infuria*), verbo + nome (*fare una passeggiata, ingranare la marcia*), nome + aggettivo (*piatto freddo, caffè macchiato*), verbo + avverbio (*pagare profumatamente, scusarsi umilmente*), nome + di + nome (*un mazzo di chiavi, un banco di pesci*). Ed anche se nelle diverse lingue troviamo le stesse categorie, le espressioni di solito non corrispondono, ma se le lingue sono vicine tipologicamente, ci possono essere somiglianze.

Dal punto di vista dell'apprendimento di una seconda lingua, le collocazioni entrano con molta difficoltà nell'interlingua degli apprendenti. E ciò accade anche quando le due lingue sono affini, ad esempio italiano e spagnolo. In effetti, la trasparenza semantica delle collocazioni, unita alla somiglianza fra le due lingue, ne facilita la comprensione e così queste passano inosservate. E nel momento della produzione, la stessa trasparenza semantica che fa scambiare le collocazioni per combinazioni comuni – prodotte cioè da regole di solidarietà lessicale - e l'esistenza nello spagnolo di collocazioni simili ma non uguali a quelle italiane fanno sì che nell'interlingua degli studenti spagnoli compaiano frasi come [1a] o [1b]:

[1a] `Le frisse nell'olio molto caldo perché` **`si fecero presto`**`_senza bruciarsi.` Studente di livello avanzato

[1b] `Affinché i soldati` **`rimanessero pieni`**`.` Studente di livello avanzato

Nel primo esempio lo studente ha tradotto letteralmente una collocazione dello spagnolo *hacerse de prisa* invece di dire *cuocere subito*. Nel secondo ha creato una combinazione originale, caratteristica dell'interlingua, fondendo due collocazioni spagnole *quedar satisfecho* [*rimanere soddisfatto*] y *sentirse lleno* [*sentirsi pieno*] al posto di *sentirsi sazio*.

---

[1]Cfr. Marello 1996 e Simone 1990.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 323-333.

Il problema dell'apprendimento delle collocazioni è aggravato dal fatto che, diversamente dalle frasi idiomatiche e dai modi di dire, di solito queste non sono messe in evidenza nei corsi di lingua, per cui gli studenti non le registrano e quindi non le assimilano come lessemi complessi, unica garanzia per poterle usare correttamente. Per di più nei dizionari se ne trovano ben poche e non ci sono ancora raccolte apposite dove poterle rintracciare.

2. COME REPERIRE LE COLLOCAZIONI NEI NUNC. Tornando al discorso delle potenzialità dei NUNC, abbiamo rilevato che questi corpora possono essere di grande aiuto per cercare di risolvere il problema dell'apprendimento delle collocazioni. Offrono infatti, come vedremo, la possibilità di trovare un gran numero di questi "pacchetti di parole" – come le chiama C. Marello – usati in contesti diversi, materiale che l'insegnante potrà usare per organizzare attività da proporre a lezione e gli studenti per cercare le combinazioni che non conosce.

Per aiutare gli studenti ad usare i NUNC ed a scoprirne l'utilità, abbiamo disegnato una serie di attività comunicative di scrittura e di conversazione rivolte ad apprendenti con diversi livelli di competenza, dal livello elementare a quello avanzato. Per ogni attività abbiamo preparato una scheda di lavoro in cui viene indicato il compito da svolgere ed i passi da fare per usare i NUNC come strumento atto a risolvere, tra l'altro, i problemi di collocazioni. Abbiamo suggerito di fare la *Ricerca linguistica* (più parole) e non la *Ricerca semplice* (una sola parola) per ragioni di coerenza con l'obiettivo di questo lavoro. Inoltre nella scheda abbiamo dato delle istruzioni su come effettuare la ricerca delle collocazioni, come alternativa a quelle proposte dagli autori dei corpora, perché abbiamo considerato che queste ultime sono estremamente complesse per i non specialisti e non del tutto consone all'obiettivo che ci siamo prefissati. Queste istruzioni riguardano sette percorsi di ricerca che rispecchiano, meno il primo, le strutture delle categorie di collocazione. Il primo percorso riguarda, infatti, la ricerca di una sola parola, il che può sembrare contraddittorio visto che il nostro scopo era aiutare gli studenti ad apprendere combinazioni di parole. Ma abbiamo pensato che non era il caso di rinunciare a questa opzione che offrono i NUNC di arrivare, partendo da una sola parola, ad una collocazione, anche se non messa in evidenza graficamente dal sistema.

2.1 PRIMO PERCORSO: UNA PAROLA. La ricerca di una sola parola può servire per verificare le ipotesi dell'alunno sul significato, sulla grammatica o sui contesti d'uso di una data parola. Ad es., se lo studente volesse controllare gli argomenti del verbo *bollire*, potrebbe:

Cliccare "inizio parola" + scrivere *bollire* nella casella in bianco accanto a "il lemma" + cliccare "il lemma"[2] + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

Tav. 1: Primo percorso: una parola.

Dai risultati della ricerca lo studente scoprirà che *bollire* è un verbo monovalente, come in [2], ed anche bivalente, come in [3a] e [3b]:

[2a] `litri versa il latte , lo zucchero e la scorza grattugiata di 1/2 limone . Metti sul fuoco e quando` **`bollirà`** `aggiungi il riso ed un pizzico di sale . Quando il riso sarà cotto ( il latte sarà stato assorbito`[3] NUNC-IT Cucina;

[2] Abbiamo suggerito di inserire la parola da cui parte la ricerca nello spazio riservato a "il lemma" invece di quello riservato a "la parola" perché con la prima opzione si ottiene un maggior numero di combinazioni.

[3] Per gli esempi tratti dai NUNC abbiamo selezionato *contesti* di 20 parole e 20 o 100 *risultati* per volta.

[3a] `Quando uso i malti per rinforzare la birra ( come fermentare al posto dello zucchero per capirci ) quanto devo` **`bollirlo`** `? 1 per il malto in polvere 15-20 minuti possono bastare 1 Spero in una pronta risposta Gianluca 1258 ansi` NUNC-IT Cucina,

[3b] `scontando lo stesso procedimento . Allora , io metto 1 Kg di miele ogni 4 Lt di acqua ; faccio` **`bollire`** `almeno un ora ( almeno il tanto che smetta di fare quella schiuma e il tanto di rimuoverla tutta` NUNC-IT Cucina.

2.2 SECONDO PERCORSO: NOME + VERBO. Questa collocazione è composta da un nome, che funge da soggetto, e da un verbo. È difficile rintracciarla nei NUNC, perchè a volte i due elementi che la costituiscono compaiono separati. Comunque ci saranno dei momenti in cui lo studente ne avrà bisogno. Ad esempio, se non sa come dire con una sola espressione che la macchina è andata fuori strada perché ha slittato, cioè che *la macchina ha sbandato*, dovrà fare i seguenti passi:

Cliccare "inizio parola" + scrivere *macchina* nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola" + scrivere *ha* nella casella in bianco accanto a "la parola" + cliccare "la parola" + cliccare "fine parola" + cliccare "un part.pass." + cliccare "dunque invia la richiesta che hai formulato"

Tav. 2: Secondo percorso: nome + verbo

Il contesto in cui compare la collocazione *la macchina ha sbandato*, esempio [4], aiuta lo studente ad identificare facilmente che quella è l'espressione che sta cercando. Infatti riguarda un incidente in cui viene descritta l'azione di *sbandare.*

[4] `serata trascorsa tra amici in un pub . All' improvviso il conducente dell' auto ha perso il controllo ,` **`la macchina ha sbandato`** `ed è finita nella corsia opposta dove sopraggiungeva un Tir . 1 Quando si leggono queste notizie viene da giustificarle .` NUNC-IT Motori.

2.3 TERZO PERCORSO: VERBO + NOME. Collocazione in cui il nome funge da complemento oggetto. La ricerca può seguire due strade. Si può iniziare dal nome. Se lo studente conosce il sostantivo *foto*, ma non sa con quali verbi si può combinare per esprimere l'idea della ripresa di un'immagine con la macchina fotografica, può fare quanto segue:

Cliccare "inizio parola" + cliccare "un infinito"+ cliccare "fine parola" + cliccare "inizio parola"+ scrivere *foto* nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

Tav. 3: Terzo percorso: verbo + nome

I risultati della ricerca, come si vede negli esempi [5a-d], danno allo studente la possibilità di individuare sia la collocazione vera e propria, l'espressione più precisa, *scattare foto,* sia altre combinazioni. Tra queste una molto frequente, *fare foto,* ed altre meno come *effettuare foto* e *creare foto*.

[5a] NEOFITA DELLA FOTOGRAFIA ... Ho provato anch' io una V3 e la qualità di immagine durante la ripresa (senza **scattare foto**) è perfetta !!! Quando scrivi foto vengono visualizzate benissimo ... intendi dire : visionando " fotografie " già scattate NUNC-IT Foto,

[5b] grande passo " , vorrei sondare un po' i pareri di altri utilizzatori di questo modello . In particolare mi interessa **fare foto** di paesaggi , quasi esclusivamente montani , spesso a quote elevate (4500 m) e a basse temperature NUNC-IT Foto,

[5c] Ciao a tutti , mi piacerebbe ricevere dei consigli su come **effettuare foto** urbane notturne sfruttando le soli luci della città ... Se fai le foto quando c' è la luna piena NUNC-IT Foto,

[5d] una grandissima attenzione perché che sia il peso molto contenuto sia un tempo di scatto non molto basso concorrono a **creare foto** mosse . Diciamo che non vorrei ripetere questa esperienza ! : -(Vi ringrazio in anticipo Argonath 18533 NUNC-IT Foto.

Si può iniziare la ricerca anche dal verbo. Se lo studente conosce, ad esempio, il verbo *sviluppare*, ma non sa a quali sostantivi lo può abbinare, dovrà fare i seguenti passi:

Cliccare "inizio parola" + scrivere *sviluppare* nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola" + cliccare "un nome" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

Tav. 4: Terzo percorso: verbo + nome.

A noi pare che questa modalità di ricerca sarà molto meno produttiva di quella precedente perché è più probabile che uno studente conosca il nome della collocazione e non il verbo.

2.4 QUARTO PERCORSO: NOME + AGGETTIVO. Anche in questo caso il nucleo della collocazione è il sostantivo, per cui proponiamo di iniziare la ricerca dal nome e non dall'aggettivo. Ad esempio se lo studente vuole indicare vino non imbottigliato, ma non sa che aggettivo usare per identificarlo, può fare la seguente strada nel NUNC-IT Cucina:

Cliccare "inizio parola" + scrivere *vino* nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola" + cliccare "un aggettivo" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

Tav. 5: Quarto percorso: nome + aggettivo.

Così verrà a sapere che in italiano il vino non imbottigliato viene denominato *vino sfuso*. Cfr. esempio [6]:

[6] solo locale è veramente da record !) x non annoiarvi vi racconto solo il primo : chiediamo mezzo litro di **vino sfuso** della casa e l' oste (il gestore) prende una bottiglia già stappata , assaggiata e rifiutata dal tavolo NUNC-IT Cucina.

Può darsi che inizialmente lo studente non capisca il significato dell'aggettivo *sfuso,* ma ci arriverà senz'altro non appena leggerà attentamente il testo, dove l'aggettivo è seguito da un sintagma preposizionale che lo spiega, *della casa.* E tradizionalmente il vino della casa servito nelle trattorie è sfuso.

2.5 QUINTO PERCORSO: VERBO + AVVERBIO. Il nucleo di questo tipo di collocazione è il verbo. Di conseguenza, è questo l'elemento da cui deve partire la ricerca. Se lo studente deve indicare, in una ricetta, il modo in cui si taglia di solito un alimento o che ha pagato una cifra esagerata per un certo servizio od il modo in cui è stato salutato da un amico potrebbe fare una ricerca di questo tipo:

Cliccare "inizio parola" + scrivere il verbo (*tagliare* o *pagare* o *salutare*) nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola" + cliccare "un avverbio" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

Tav. 6: Quinto percorso: verbo + avverbio.

Dai risultati della ricerca si evince che gli avverbi che formano parte di questa categoria di collocazione sono di modo [7a-c] e di intensità [8].

[7a] `2 cucchiai di basilico fresco , sminuzzato rametti freschi di timo , per guarnire 1 1. pelare i pomodori e` **`tagliarli grossolanamente`** `2. scaldare l' olio in una padella e fare saltare per 3 minuti cipolla e aglio 3. nel frattempo scaldare` NUNC-IT Cucina,

[7b] `pesce spada affumicato, ottima alternativa all' inflazionato salmone ; della bottarga di muggine non troppo stagionata che puoi servire` **`tagliata sottilmente`** `su un letto di sedano ( tagliato a piccoli pezzi ) con un filo d' olio d' oliva extravergine accompagnata` NUNC-IT Cucina,

[7c] `alcolico rappresentava una controindicazione ! Secondo . I vini sono senz' altro interessanti ma la E&J Gallo se li fa` **`pagare profumatamente`** `. Il marchio e la struttura della grande azienda si fanno pagare . Terzo . La storia di Ernest and` NUNC-IT Cucina;

[8] `entrai incuriosito perchè proponevano il test per la strada , lo feci , rimasi un po' perplesso , li` **`salutai caramente`** `convinto che fossero una marea di squallidi esaltati . Ogni volta che ripassavo per la via per un certo periodo` NUNC-IT Generico.

2.6 SESTO PERCORSO: NOME + DI + NOME. Le collocazioni di questo tipo indicano l'unità di cui forma parte un'entità più piccola oppure il gruppo a cui appartiene un certo individuo, ad esempio *spicchio d'aglio* o *gregge di pecore*. Il primo nome della combinazione indica il gruppo o l'unità a cui appartiene l'individuo o l'entità indicati dal secondo nome[4]. A noi pare che la ricerca in questo caso dovrebbe cominciare dal secondo nome perché i nomi che occupano la prima posizione nella collocazione sono dei quantificatori che compaiono con un numero molto limitato di nomi quindi molto probabilmente gli studenti non li conoscono.

[4] Cfr. Corpas Pastor 1996, p. 74

Se lo studente ha bisogno di reperire una collocazione di questo tipo, ad esempio per parlare di una porzione di pane o di aglio, suggeriamo il seguente percorso (di cui cfr. gli ess. [9a-d]):

Cliccare "inizio parola" + cliccare "un nome" + cliccare "fine parola" + cliccare "inizio parola" + scrivere *di* nella casella in bianco accanto a "la parola" + cliccare "la parola" + cliccare "fine parola" + cliccare "inizio parola" + scrivere il nome corrispondente (*pane, aglio*) nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

Tav. 7: Sesto percorso: nome + di + nome.

[9a] , in modo che siano sopra morbide e sotto dorate . Nel frattempo fate scaldare nel tostapane una o **due fette di pane** con la mollica ben compatta. azz .... e meno male che è uno spuntino dietetico !!!!!! :O(((NUNC-IT Cucina,

[9b] acciughe tritate 1 C capperi 2 rossi d' uovo 1 C olio sale e pepe , cognac 1 svuotare **il filone di pane** . mescolare tutti gli ingredienti e insaporire con sale pepe e cognac . riempire il pane avvolgere nella carta stagnola NUNC-IT Cucina,

[9c] (colmo) di parmigiano grattugiato sale pepe 2 PREPARAZIONE Scaldare l' olio in una larga padella . Rosolatevi **gli spicchi di aglio** mondati ; schiacciateli con una forchetta , mentre vanno prendendo colore ; poi levateli dal recipiente ed eliminateli . NUNC-IT Cucina,

[9d] risultato finale ... per esempio, mo' - se ci avessi delle bietoline - le stuferei cinque minuti con **una puntina di aglio** e peperoncino 1 soffritto o no ? 1 - e ci farei sciogliere pure un' alicetta ... tié! Poi NUNC-IT Generico.

Questi risultati potrebbero spingere lo studente ad allargare la ricerca, a domandarsi con quali altri nomi si possono combinare le parole *fetta* e *spicchio*, inizando così una nuova ricerca che gli permetterebe di scoprire, per esempio, che si può parlare anche di *una fetta di prosciutto, di pollo* e *di torta* e di *uno spicchio di limone* o *di pera*.

3. RISULTATI DELLE ATTIVITÀ. Le attività sono state testate a lezione – nella sala computer – con diversi gruppi di studenti adulti.[5] Prima di iniziare ogni attività l'insegnante ha letto con gli studenti la scheda di lavoro per assicurarsi che questi avessero capito cosa *dovevano* e cosa *potevano* fare. Era importante, infatti, che non scambiassero l'obiettivo finale, la realizzazione del compito (il dialogo, la scrittura di un e-mail, ecc.), con la ricerca nei corpora, un mezzo, quest'ultimo, per soddisfare le necessità personali di ogni singolo utente impegnato nel raggiungimento dell'obiettivo. Inoltre, l'insegnante ha spiegato agli studenti che, oltre ai quattro corpora elettronici, potevano consultare anche un dizionario nel caso in cui non sapessero da dove iniziare la ricerca.

La maggioranza degli studenti ha svolto le attività senza necessità di ulteriori chiarimenti da parte dell'insegnante, e solo le persone meno abituate ad usare il computer hanno avuto bisogno di aiuto per le prime due o tre ricerche. Ciò dimostra che le istruzioni di ricerca inserite nelle

[5] Gli studenti erano di diversi livelli di competenza – elementare, intermedio, avanzato – ed iscritti ai corsi di italiano delle *Escuelas Oficiales de Idiomas* di Madrid (due sedi) e di Segovia nell'anno 2005. Le *Escuelas Oficiales de Idiomas* sono scuole statali che offrono corsi di lingue ad adulti (dai 16 anni in avanti).

schede di lavoro erano valide. Invece il tempo programmato per ogni attività, due ore per non andare oltre la durata della lezione, a volte si è dimostrato insufficiente perché gli studenti, incuriositi dal nuovo sussidio didattico, si sono buttati a capofitto nella ricerca di combinazioni di parole dimenticando l'obiettivo finale.

Comunque tutti gli studenti sono riusciti a portare a termine il lavoro, con buoni risultati, con entusiasmo ed in modo autonomo. Il che conferma che i NUNC sono un ottimo strumento per l'apprendimento dell'italiano come lingua straniera soprattutto perché stimola gli studenti a diventare autonomi e li aiuta a tagliare il cordone ombelicale che li mantiene legati all'insegnante.

4. APPENDICE.

4.1 ATTIVITÀ 1: IN UN DISTRIBUTORE DI BENZINA.

LIVELLO: Elementare

COMPITO[6]: Dialogo fra due studenti, uno nella parte del benzinaio e l'altro in quella del cliente.

CLIENTE: Stai girando l'Italia in macchina. Ti fermi ad un distributore di benzina perché sei in riserva. Rivolgiti al benzinaio che si sta avvicinando.

BENZINAIO: Una macchina si è fermata ad una pompa del tuo distributore. Avvicinati e servi il cliente.

PAROLE / ESPRESSIONI PER SVOLGERE IL COMPITO: Per svolgere il compito puoi usare le seguenti parole / espressioni, oltre a quelle che già conosci.

PER ORDINARE IL CARBURANTE

Il pieno (di benzina, di diesel)
Benzina verde / senza piombo
Gasolio / Diesel
30 €

ALTRI SERVIZI

Controllare l'acqua, l'olio
Controllare le gomme / la pressione delle gomme

PER PAGARE

Pagare con la carta di credito
Pagare in contanti
Accettare la carta di credito

4.2 ATTIVITÀ 2: UNA RICETTA.

LIVELLO: Intermedio

COMPITO: Un tuo amico ti ha chiesto la ricetta di un piatto che gli hai preparato l'ultima volta che è stato a casa tua. Mandagliela via e-mail.

ISTRUZIONI PER L'USO DEI MATERIALI DI CONSULTAZIONE: Se non sei sicuro o non conosci tutte le parole/espressioni per scrivere la ricetta, oltre ai vocabolari, puoi usare il NUNC generico ed il NUNC cucina (www.corpora.unito.it).

SUGGERIMENTI PER USARE I NUNC nel modo più semplice ed efficace.

- Entra nella pagina web www.corpora.unito.it
- Clicca NUNC generico o NUNC cucina.
- Clicca Ricerca linguistica.

[6] Ad ogni studente viene consegnata una scheda di lavoro che contiene solo le istruzioni che lo riguardano.

- In questa pagina c'è la possibilità di cliccare Istruzioni per l'uso. Ma, per il tipo di lavoro che devi fare, noi ti proponiamo i seguenti percorsi:
  1. Se cerchi una sola parola:

Cliccare "inizio parola" + scrivere la parola corrispondente nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

  2 Se cerchi una combinazione di nome + verbo:

Cliccare "inizio parola" + scrivere il nome nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola" + cliccare "un verbo" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

  3 Se cerchi una combinazione di verbo + nome (dato):

Cliccare "inizio parola" + cliccare "un infinito"+ cliccare "fine parola" + cliccare "inizio parola"+ scrivere il nome nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

  4 Se cerchi una combinazione di verbo (dato) + nome:

Cliccare "inizio parola" +scrivere il verbo nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola"+ cliccare "un nome" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

  5 Se cerchi una combinazione di nome + aggettivo:

Cliccare "inizio parola" + scrivere il nome nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola" + cliccare "un aggettivo" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

  6 Se cerchi una combinazione di verbo + avverbio:

Cliccare "inizio parola" + scrivere il verbo nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "inizio parola" + cliccare "un avverbio" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

  7 Se cerchi una combinazione di nome + di + nome (dato):

Cliccare "inizio parola" + cliccare "un nome" + cliccare "fine parola" + cliccare "inizio parola" + scrivere *di* nella casella in bianco accanto a "la parola" + cliccare "la parola" + cliccare "fine parola" + cliccare "inizio parola" + scrivere il nome nella casella in bianco accanto a "il lemma" + cliccare "il lemma" + cliccare "fine parola" + cliccare "dunque invia la richiesta che hai formulato"

4.3 ATTIVITÀ 3: CONSIGLIARE UN RISTORANTE.

LIVELLO: Intermedio.

COMPITO: Dialogo fra due studenti, nella parte di due amici.

STUDENTE A: sei in partenza per l'Italia e ti fermerai per qualche giorno a Napoli, un posto che il tuo amico conosce bene. Chiedigli consiglio su qualche locale in cui si possano assaggiare specialità italiane o napoletane e dove si mangia bene e senza spendere molto.

STUDENTE B: Sei un buongustaio e conosci bene Napoli. Un tuo amico ci andrà qualche giorno e ti chiede di parlargli di ristoranti e di piatti tipici. I locali possono essere reali od immaginari. Prova ad aiutarlo.

ISTRUZIONI PER SVOLGERE IL COMPITO: A *Fase di preparazione* (a casa)

A1. Pensa alle parole / espressioni di cui hai bisogno per svolgere il compito nel ruolo che ti è stato assegnato (Studente A o Studente B). Per cercare quelle che ti mancano, oltre ai vocabolari, puoi usare il NUNC generico ed il NUNC cucina www.corpora.unito.it (vedi sotto).

A2. Per informazioni sui piatti della cucina italiana usa un libro di cucina od internet.

ISTRUZIONI PER SVOLGERE IL COMPITO: B *Dialogo* (in classe)

SUGGERIMENTI PER USARE I NUNC nel modo più semplice ed efficace.

- Entra nella pagina web www.corpora.unito.it
- Clicca NUNC generico o NUNC cucina.
- Clicca Ricerca linguistica.
- In questa pagina c'è la possibilità di cliccare Istruzioni per l'uso. Ma, per il tipo di lavoro che devi fare, noi ti proponiamo i seguenti percorsi. (Vedi attività 2)

4.4 ATTIVITÀ 4: DESCRIVERE UNA MACCHINA FOTOGRAFICA.

LIVELLO: Avanzato

COMPITO: Dialogo fra due studenti, nella parte di due amici.

STUDENTE A: Vuoi comprare una buona macchina fotografica perché ti stai appassionando alla fotografia. Non sai che modello comprare ed allora chiedi aiuto ad un amico.

STUDENTE B: Sei un amante della fotografia ed un tuo amico, che vorrebbe comprare una macchina fotografica, ti chiede un consiglio. Dagli una mano.

ISTRUZIONI PER SVOLGERE IL COMPITO: A *Fase di preparazione* (a casa)

Pensa alle parole / espressioni di cui hai bisogno per svolgere il compito nel ruolo che ti è stato assegnato (Studente A o Studente B). Per cercare quelle che ti mancano, oltre ai vocabolari, puoi usare il NUNC generico ed il NUNC fotografia www.corpora.unito.it (vedi sotto).

ISTRUZIONI PER SVOLGERE IL COMPITO: B *Dialogo* (in classe)

SUGGERIMENTI PER USARE I NUNC nel modo più semplice ed efficace.

- Entra nella pagina web www.corpora.unito.it
- Clicca NUNC generico o NUNC fotografia.
- Clicca Ricerca linguistica.
- In questa pagina c'è la possibilità di cliccare Istruzioni per l'uso. Ma, per il tipo di lavoro che devi fare, noi ti proponiamo i seguenti percorsi. (Vedi attività 2)

4.5 ATTIVITÀ 5: CON CHE MACCHINA ANDIAMO?

LIVELLO: Avanzato.

COMPITO: Discussione fra tre studenti, nella parte di tre amici.

STUDENTE A: Tu ed altri due amici state organizzando un lungo viaggio in macchina, sistemazione campeggio. Dovete decidere con quale macchina andare. Metti la tua a disposizione. Hai una Fiat Stilo 3p.

STUDENTE B: Tu ed altri due amici state organizzando un lungo viaggio in macchina, sistemazione campeggio. Dovete decidere con quale macchina andare. Metti la tua a disposizione. Hai una Seat Alhambra.

STUDENTE C: Tu ed altri due amici state organizzando un lungo viaggio in macchina, sistemazione campeggio. Dovete decidere con quale macchina andare. Avresti la possibilità di avere in prestito un vecchio camper Ducato Adria.

ISTRUZIONI PER SVOLGERE L'ATTIVITÀ: A *Fase di preparazione* (a casa)

Se non conosci le caratteristiche essenziali del modello che ti è stato assegnato, fai una ricerca sul NUNC motori o su Internet. Pensa poi alle parole / espressioni di cui hai bisogno per parlare dei vantaggi e degli svantaggi di ogni tipo di macchina. Per cercare quelle che ti mancano, oltre ai vocabolari, puoi usare il NUNC generico ed il NUNC motori www.corpora.unito.it.

ISTRUZIONI PER SVOLGERE L'ATTIVITÀ: B *Discussione* allo scopo di stabilire quale dei tre mezzi sia il più adeguato alle vostre esigenze (in classe).

SUGGERIMENTI PER USARE I NUNC nel modo più semplice ed efficace.

- Entra nella pagina web www.corpora.unito.it
- Clicca NUNC generico o NUNC motori.
- Clicca Ricerca linguistica.
- In questa pagina c'è la possibilità di cliccare Istruzioni per l'uso. Ma, per il tipo di lavoro che devi fare, noi ti proponiamo i seguenti percorsi. (Vedi attività 2)

BIBLIOGRAFIA.

BIBER - CONRAD- REPPEN

2000 Douglas Biber - Susan Conrad - Randi Reppen, *Corpus Linguistics. Investigating language structure and use,* Cambridge, Cambridge University Press, 2000.

BOSQUE el alii

2004 *Redes: Diccionario combinatorio del español contemporáneo*, a cura di Ignacio Bosque, Madrid, SM, 2004.

CARTER - MCCARTHY

2006 Ronald Carter - Michael McCarthy, *Cambridge Grammar of English. A Comprehensive Guide. Spoken and Written English Grammar and Usage,* Cambridge, Cambridge University Press, 2006.

CORINO

¶ 13. Elisa Corino, *NUNC (Newsgroup UseNet Corpora). Aspetti della testualità e questioni metodologiche*, in questo volume, pp. 225-252.

CORPAS

1996 Gloria Corpas Pastor, *Manual de Fraseología Española*, Madrid, Editorial Gredos, 1996.

DE BEAUGRANDE

2002 Robert De Beaugrande, *Descriptive Linguistics at the Millennium: Corpus Data as Authentic Language,* in "Journal of Language and Linguistics" I (2002)[2] 91-131, disponibile online alla pagina `http://www.shakespeare.uk.net/journal/vol1no2_home.html`.

MARELLO

1996 Carla Marello, *Le parole dell'italiano. Lessico e Dizionari*, Bologna, Zanichelli, 1996.

MCCARTHY

1991 Michael McCarthy, *Discourse Analysis for Language Teachers*, Cambridge, Cambridge University Press, 1991.

ORTEGA
i.s. Ángeles Ortega Calvo, *Investigaciones recientes en lingüística y sus implicaciones para la enseñanza de idiomas,* Madrid, Ministerio de Educación y Ciencia, in corso di stampa in *Las nuevas enseñanzas de las escuelas oficiales de idiomas: renovación metodológica*, Instituto Superior de Formación del Profesorado, "Colección Aulas de verano - Serie Humanidades".

SIMONE
1990 Raffaele Simone, *Fondamenti di linguistica,* Roma-Bari, Editori Laterza, 1990.

CORPORA DI RIFERIMENTO.

Corpora.unito.it	`http://www.corpora.unito.it/`.
NUNC-IT Generic I	`http://www.bmanuel.org/projects/ng-HOME.html`.
NUNC-IT Cooking	`http://www.bmanuel.org/projects/ng-HOME.html`.
NUNC-IT Photo	`http://www.bmanuel.org/projects/ng-HOME.html`.
NUNC-IT Motor	`http://www.bmanuel.org/projects/ng-HOME.html`.

JACQUELINE VISCONTI
(Università di Genova)

20. Corpora ed analisi testuali. *La particella* mica.

0. PREMESSA. L'uso di corpora come fonti di dati "reali" apre nuovi orizzonti all'analisi linguistica. La possibilità di fondare l'indagine su basi quantitative facilita l'elaborazione di ipotesi in certa misura "oggettive" o quantomeno falsificabili. La creazione e diffusione di corpora ad opera del gruppo di ricerca torinese, coordinato da C. Marello e M. Barbera, colma in particolare una lacuna importante: la carenza di corpora di qualità liberamente disponibili per la lingua italiana (cfr. Barbera ¶ 1 in questo volume). Uno dei pregi della raccolta torinese (`www.corpora.unito.it`) è la costituzione di corpora esplicitamente finalizzati a ricerche di tipo testuale. In questo contributo porrò in evidenza la "versatilità testuale" di uno di questi, il *Newsgroups UseNet Corpus* (NUNC), grazie all'analisi della particella *mica*, la cui semantica dipende fortemente da restrizioni co-testuali.

1. MICA[1]. L'uso di *mica* è tradizionalmente associato ad una funzione "enfatica", o "presupposizionale": per Bernini - Ramat 1992, pp. 25-26, l'impiego di *mica* «implica che il parlante presuppone che quanto egli nega sia invece ritenuto vero od atteso come realizzabile dal suo interlocutore»; per Manzotti - Rigamonti 1991, p. 284, «*mica* non nega una asserzione, ma una presupposizione di quella asserzione. Così una frase come: 'Non fa *mica* freddo fuori' è la replica adeguata a: 'Mettiti la sciarpa quando esci', che presuppone: 'Fa freddo fuori', non ad una domanda come: 'Fa freddo fuori?'»; per Zanuttini 1997, p. 61 (cfr. analogamente Cinque 1976/91) «the occurrence of *mica* is pragmatically restricted to those contexts in which the non-negative counterpart of the proposition expressed by the sentence is assumed in the discourse. For example, in order for *mica* to be uttered felicitously in: 'Gianni non ha [*mica*] la macchina', it is necessary that the proposition that Gianni has a car be entailed by the common ground. If such a proposition is not part of the common ground, the presence of *mica* renders the sentence infelicitous and its counterpart without *mica* must be used».

[1a] A. Chi viene a prenderti?
B. Non so. Ma Gianni non ha (**mica*) la macchina[2].

[1b] A. Chi viene a prenderti, Gianni?
B. Non so. Ma Gianni non ha (*mica*) la macchina.

Definizioni di questo tipo, in termini di enfasi o presupposizioni, sono oggetto di critica in Schwenter *i.s.* (e Schwenter 2003, p. 1001), che nota come questi concetti siano raramente definiti con chiarezza. Anche Zanuttini, nota Schwenter, non chiarisce il suo uso di *common ground*; anzi, la sua definizione, se per *common ground* si intende, *à la* Stalnaker, l'insieme di proposizioni condivise e considerate come vere dagli interlocutori, non rende conto del fatto che l'esempio [1a] non sia accettabile neanche in un contesto in cui gli interlocutori dividano un *common ground* in cui Gianni ha la macchina, è solito venire a prendere B, ecc.

---

[1] Una versione più estesa di questa sezione è in Visconti *i.s.*

[2] Salvo diversamente avvisato, gli esempi in Courier sono tratti dal Corpus NUNC-IT Generic I; in Times, invece, sono gli *exempla ficta* e quelli tratti da altre fonti.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 335-345.

Per affinare la caratterizzazione di questo tipo di negazione, "non canonica", Schwenter 2003 fa riferimento alla nozione informativa di "accessibilità": la proposizione negata da *non V mica* deve essere accessibile, direttamente o tramite un'inferenza, dal contesto discorsivo in cui occorre la negazione. Schwenter *i.s.* spinge oltre questa intuizione, argomentando come le condizioni di impiego della negazione non canonica in catalano, italiano e portoghese brasiliano rilevino della struttura informativa del discorso, coincidendo con la negazione di una proposizione *discourse-old* (nei termini di Prince 1992) e *salient*, od *activated* (nel senso di Dryer 1996).

Consideriamo più da vicino questi concetti, la cui origine, com'è noto, è già nei lavori di Chafe 1976 sulla struttura informativa del discorso. Per Chafe, la distinzione tra informazione data e nuova è in termini dello statuto cognitivo dei referenti interessati; in particolare, è informazione data: «that knowledge which the speaker assumes to be in the consciousness of the addressee at the time of the utterance» (Chafe 1976, p. 30); nuova: «what the speaker assumes he is introducing into the addressee's consciousness by what he says» (*ib.*). Riprendendo questa concezione, Prince distingue esplicitamente tra datità nel senso di *salienza*: «The speaker assumes that the hearer has or could appropriately have some particular thing/entity etc. in his/her consciousness at the time of hearing the utterance» (Prince 1981, p. 228), la concezione di Chafe; e datità nel senso di *shared knowledge*: «The speaker assumes that the hearer 'knows', assumes, or can infer a particular thing (but it's not necessarily thinking about it)» (*ib.*, p. 230); «information the speaker believes the listener already knows and accepts as true» (*ib.*, p. 231). Così Dryer 1996 distingue tra dato nel senso di *presupposto* (presupposizione pragmatica): «part of the *common ground*, the set of propositions that the speaker believes and assumes the hearer to believe» (Stalnaker 1974, p. 199), e dato nel senso di *attivato*, presente all'attenzione dell'interlocutore in un certo istante (statuto cognitivo). Dryer individua inoltre entità *accessibili*: «related by inference or other type of association to an activated entity, thus highly accessible to activation, as in 'John came into the room with a woman we had never met. We wondered where *his wife* was'» (Dryer 1996, p. 519); per concludere: «There are activated beliefs, nonactivated beliefs and also activated propositions that are not believed» (*ib.*).

Tale opposizione in termini di proprietà cognitive si intreccia con la concezione classica della datità, circa l'introduzione esplicita o meno in un certo mondo di riferimento, nella distinzione di Prince 1981 tra *nuovo*: «when a speaker first introduces an entity into the discourse» (*ib.*, p. 235),

[2] I bought a beautiful dress;

evocato: «referring to an entity already in the discourse-model» (*ib.*, p. 236), come in:

[3] Susie went to visit her grandmother and *the sweet lady* was making Peking Duck;

ed *inferibile*: «if the speaker assumes the hearer can infer it, via logical – or, more commonly, plausible – reasoning, from discourse entities already Evoked or from other Inferrables» (*ib.*, p. 236), come in:

[4] I went to the post office and *the stupid clerk* couldn't find a stamp.

Si noti che, per Chafe, un sintagma nominale è dato se il suo referente è stato «explicitly introduced in the discourse or be present in the physical context or be categorized in the same way as a referent previously introduced or physically present» (Chafe 1976, p. 32). Così, *the beer* è dato in [5a] e nuovo in [5b]:

[5a] We got *some beer* out of the trunk. *The beer* was warm.
[5b] We got the *picnic supplies* out of the trunk. *The beer* was warm.

La questione degli inferibili introduce un affascinante elemento di complicazione: sono *nuovi* (non erano prima nell'universo discorsivo) o *dati* (in quanto elaborati a partire da entità già dell'universo di discorso)? In un recente contributo sull'argomento, Birner 2006, p. 15, riprende la classificazione di Prince 1992:

	Hearer-old:	Hearer-new:
Discourse-old:	Previously evoked	(Non-occurring)
Discourse-new:	Not evoked, but known	Brand-new

Tav. 1: La classificazione di Prince 1992.

e propone di ridefinire la nozione di *discourse-old* in termini di inferenze, non di menzione esplicita. Secondo l'autrice: «it is the presence of inferential link, not explicit prior evocation, that defines the class of information treated as discourse-old. In the case of explicitly evoked information, the inferential relation is identity» (Birner 2006, p. 20): tale è la relazione tra *her grandmother* e *the sweet lady* nell'esempio [3], o tra *some beer* e *the beer* nell'esempio [5a]. Riferendosi alla letteratura psicolinguistica, Birner distingue in particolare due tipi di inferenze: (j) "forward", od "elaborating"; (ij) "backward", o "bridging" (*ib.*, pp. 23-24). Le prime sono immediatamente provocate da un "trigger", come *the post office* → *clerk* nell'esempio [4], o *get married* → *wedding* in [6]:

[6] She *got married* recently and at the *wedding* was the mother, the stepmother and Debbie.

Le seconde, invece, non sono tratte se non *a posteriori*, quando si renda necessario stabilire coerenza tra un segmento di discorso ed il discorso precedente, come nell'esempio [5b], in cui *picnic supplies* non dà immediatamente luogo all'inferenza: *beer*. Mentre le prime sono considerate informazione nota all'interlocutore, alla stregua delle inferenze di identità, le seconde non lo sono. In uno schema (Birner 2006, p. 25):

	H-old:	*H-new:*
D-old	Evoked: Identity/Elaborating Inferrable (inferentially linked and known to hearer)	Bridging Inferrable (inferentially linked but not known to hearer)
D-new	Unused (not inferentially linked, but known to hearer)	Brand-new (not inferentially linked and not known to hearer)

Tav. 2: La classificazione di Birner 2006, p. 25.

Queste nozioni si rivelano basilari nella definizione delle condizioni d'uso di *mica.*

2. Il corpus. Il corpus, italiano, di cui mi avvalgo è il *NUNC-IT Generic I*, parte di una collezione multilingue di corpora di lingua contemporanea, tanto generici quanto specialistici[3], basati sui messaggi dei newsgroup. Come nota Barbera ¶ 1, § 2.2.5, cit. (e cfr. anche Corino ¶ 13), cui si rimanda per una trattazione di vantaggi e svantaggi di tale base testuale: «Un newsgroup è un forum telematico a libero accesso, gratuito, disponibile su Internet, che si manifesta nella forma di testi scritti, ed il cui funzionamento è assai semplice: ogni utente scrive un messaggio, il post, e lo invia ad una specie di "bacheca elettronica" mantenuta presso una rete di

[3] Nei settori dell'alimentazione, della fotografia e dei motori.

server (i newsserver che costituiscono Usenet), dai quali gli altri utenti del gruppo possono scaricarlo, leggerlo e rispondervi. [...] La facilità d'uso garantisce la grande diffusione dello strumento tra le categorie più diverse di utenti e giustifica la grande quantità di traffico esistente su UseNet. Queste "bacheche elettroniche" che sono i newsgroup sono poi articolate in una tassonomia precisa, ossia in un sistema di cornici argomentative che si chiamano "gerarchie", a base geografico-nazionale e/o tematica; anche queste gerarchie, peraltro, nascono dal basso in base alla iniziativa degli utenti».

Il grande interesse di questa base testuale per la nostra ricerca è il carattere fortemente "tendenziale" della varietà di lingua usata[4]. Tale tratto ci permette di cogliere quasi in tempo "reale" tendenze recenti nell'evoluzione dei costrutti studiati, come quella dell'uso di *mica non* accompagnato dalla negazione *non*. Distinti 39/50 casi di negazione frasale da 11/50 casi di negazione di costituente, ess. [7]-[9], del tipo

mica [SN]
[7a] io ho un k-2 *mica* chissà quale pc di ultima generazione
[7b] La religione dev' essere un fatto privato , *mica* una vergogna
mica [SAvv]
[8a] e un cast *mica* male
[8b] Una (*mica* tanto) breve introduzione a LaTex è reperibile
mica [SP]
[9a] siamo in India *mica* nell' obesa Italia
[9b] E parliamo dell' org di Milano , *mica* di quella di Borgonovo val di Taro

le forme rilevate per i casi di negazione frasale sono quelle riassunte nella Tav. 3 e rappresentate dagli esempi [10]-[12] seguenti:

(j)	*non... mica* (in VP)	23/39
(ij)	*mica* VP	15/39
(iij)	*mica* (in VP)	1/39

Tav. 3: *Mica* negazione frasale in NUNC-IT Generic I.

(j)
[10] Quanto affermi nei vangeli non c' è *mica* scritto !
[10] Non siamo *mica* gli americani
(ij)
[11] io *mica* ho segnalato tutti i dischi che ho , sennò ci stavo anni
[11] Ma vabbè , *mica* c' è scritto che la gente deve scrivere soltanto su certe cose
(iij)
[12] ma è *mica* colpa del Sony

Nella caratterizzazione dei contesti discorsivi di *mica*, alla luce dei concetti sopra delineati, emerge la dipendenza di *mica* da precise restrizioni discorsive, in particolare, la sua relazione con elementi dati, od "attivi" del co-testo precedente. Utile punto di partenza per l'elaborazione di una tipologia di relazioni è uno studio su *mica* nell'italiano delle origini (Visconti *i.s.*; Hansen - Visconti *i.s.*), in cui la classificazione dei possibili legami della proposizione contenente la particella con il co-testo precedente identifica quattro categorie principali: **(j)** la negazione più o

[4] Barbera ¶ 1, § 2.2.5, parla di «caratteristiche di *Umgangssprache* contemporanea»: cioè «di una lingua comune, usuale e media, non tematicamente o sociologicamente delimitabile, più vicina al parlato ma di fatto scritta, e per la quale, in realtà la dicotomia scritto-parlato non è realmente pertinente».

meno diretta di parte del co-testo precedente; **(ij)** la negazione di una presupposizione del co-testo precedente; **(iij)** la negazione di una inferenza sollecitata/resa possibile dal co-testo precedente, anche solo come attese di uno scenario, come in [15b], in cui l'inferenza negata riguarda lo scenario in cui ad un messaggero si dia risposta; **(iiij)** la ripetizione/parafrasi di parte del co-testo precedente:

(j)

[13a] E poi li disse: "Siri, se Dio vi salvi, che v'è aviso di me? Sono io ora quello T., che voi solete tanto dottare? Non vero, collui non sono *mica* *Tristano Ricc.* XIII (tosc.), App., 395 [OVI],

[13b] E allora disse lo ree: – E dunque volevi tue uccider mee overo Tristano? – Ed ella disse ke no lo vollea fare, *né mica* uccidere lui. - E dunqua volei tue uccidere pur Tristano? - Ed ella disse allora ke pur per lui l'avea fatto *Tristano Ricc.*, Cap. 3 [LIZ, XIII];

(ij)

[14a] Paura dice: "Quello omo ave molto grande avere". Sicurtade risponde: "Ciò non è *mica* omo, ma è uno grido pieno di voci" *Trattato di virtù morali*, XIII/XIV (tosc.), 25.67 [OVI],

[14b] Io sono quelo maestro per cui tutti i tereni maestri sanno tanto di bene com'egl'àno apreso; né maestri no son eglino *mica*, ché neuno no puot'esere maestro se non quegli che sa tutte le scienze *Storia SanGradale*, XIV (fior.), cap. 2, 7.18 [OVI];

(iij)

[15a] ma, se molte genti signoreggiano, con tutto che ciascuno intenda alla sua propia utilità, tuttavia ellino non sono *né mica* sì da lunga dal bene comune, come un solo, quand'elli intende al suo propio bene *Egidio Romano volg.*, 1288 (sen.), 3.2.4. [OVI],

[15b] E lo messaio trovao Tarquinio sedere in uno orto fiorito con uno bastone in mano e *mica* no li rispose, ma lo bastone ferio per li arbori e li fiori ne iectao *St. de Troia e de Roma Amb.*, 1252/58 (rom.> tosc.), 103 [OVI];

(iiij)

[16a] Allora disse la reina Isotta: - Io nol credo ke·ttue fossi figliuolo de·rree Pellinor, perké lo ree Pillinor si fue uno kortesisimo cavaliere, ma·ttue non ritrai da·ssuo legnaggio di kortesia. Impercioe ke mee non pare ke·ttue sii *mica* kortese cavaliere, quando tue davanti a mee tu mi die villania *Tristano Ricc.*, Cap. 75 [OVI],

[16b] Andò pronta et ardita, no impagorenno *mica* Buccio di Ranallo, *S. Caterina*, 1330 (aquil.) 378 [OVI].

Oltre a queste quattro tipologie principali, vi sono anche esempi (cfr. ess. [17a-d]) in cui la proposizione contenente *mica* è in una più generica relazione con il co-testo precedente:

[17a] lo vino amarostico lo corpo no notrica, la natura refutalo, no se -nde adolca *mica*; vino che ave orribile odore per certo genera in testa dolore *Regimen Sanitatis*, XIII (napol.), 576 [OVI],

[17b] Però, Amor, valer ciò mi dovrebbe; ché cchi non pecca, parmi, assai si svolpa, né non dovria portar pena *né-mica* Amico di Dante, XIII (fior.), 36.754 [OVI],

[17c] disse messer Hestor "chi puote dimorare in quella torre, che tanto è ritta per sembianti?". "Certo" disse messer T. "non vi dimora persona, se ciò non è di novello, ch'elli non è *mica* grande tempo che 'l cavaliere che manteneva quella torre fu ucciso. Ed al tempo ch'elli era vivo, dico io bene ch'elli non era in nulla terra uno passaggio sì folle come era questo" *Tristano Ricc*., XIII (tosc.), 379 [OVI],

[17d] Iohanni mio nipote sento de chiamare. Chiaschuno de vuj è parente mio carnali: vui con meco demordete e stagate, et lu meo corpu *mica* no lassète" *Legg. Transito della Madonna*, XIV (abruzz.), 26 [OVI].

Anche nei dati del NUNC, la relazione principale è quella in cui *mica* nega un'inferenza sollecitata dal co-testo precedente (15/39 casi), come negli esempi [18]:

[18a] certo che ti faccio la fattura ... ho la partita iva io ... non lavoro *mica* in nero come la maggior parte degli italiani ! ...

[18b] Al termine della discussione il professore si rivolge al primo studente : " Lei è preparato e mi piace la sua esposizione .

Approvato con – ! " . Poi si rivolge al secondo studente : " Lei ha ancora qualche incertezza , ma mi pare abbastanza preparato . Approvato con – ! " . Al terzo studente : " Lei mi ha fatto scena muta , ragazzo mio ! Più di – non posso proprio darle !!! " . E lo studente : " Ma guardi che io non debbo *mica* fare l' esame , sono solo venuto a vedere come andava l' appello ad un amico !! "

In alcuni casi (5/39), come nei testi antichi, *mica* rende esplicita un'inferenza indotta dal cotesto precedente:

[19a] Questo perché , imho , la preparazione universitaria (e scolastica , più in generale) non si è mai basata su un percorso di studio atto a preparare la persona alle situazioni " reali " , alla vita di tutti i giorni . Imho non si tratta di un modo " nuovo " di lavorare , ma casomai di apprendere con l' esperienza (e come altro che con l' esperienza ? *mica* sarà per nulla che arrivi a dirigere uno scavo quando già ne hai fatti altri come assistente , e naturalmente vale per tutti i lavori) a lavorare presto e bene e ad applicare in pratica le basi che hai acquisito nel tuo percorso di formazione .

[19b] Anche oggi per l' ennesima volta è andato in onda il solito scempio su Sky sport . Mi riferisco alla grafica che indica punteggio e tempo su Sky Sport e sky sport . Tale grafica è realmente troppo grande , troppo staccata dal margine e a peggiorare le cose si aggiungono pure le ridicole scritte accessorie : C' era così bisogno di aggiungere la scritta recupero quando l'indicatore segna il tempo oltre il esimo ? Non siamo *mica* rincoglioniti ! Lo sappiamo benissimo che ogni tempo dura minuti .

La casistica di esempi riscontrata nel NUNC induce tuttavia ad affinare la griglia interpretativa delle possibili relazioni che legano un elemento linguistico al testo adiacente. Componenti semantiche e pragmatiche si intrecciano nel definire diversi tipi di inferenze: **(a)** inferenze di tipo linguistico/presupposizionale, come una presupposizione esistenziale [20a], o comunque ancorate in un lessema – *sentirci* [20b], *tre* [20c] –, od in cui si nega e precisa il referente di un pronome [20d],

[20a] C' è però un fatto su cui dobbiamo riflettere : gli alunni ne sanno più di noi sull' utilizzo delle nuove tecnologie . Forse sarebbe bene aggiornarsi in questo campo perché è sempre utile . Ma senza assumere toni crepuscolari o apocalittici . Sì certo , loro ci sanno fare più di te con questi aggeggi maledetti , tuttavia questa curiosa situazione può offrire nuove opportunità didattiche . Ma senza farti venire i sensi di colpa : non è *mica* colpa tua se mentre loro giocavano alla playstation tu affondavi nelle griglie di valutazione !

[20b] Se scriverete un messaggio in maiuscolo , il minimo che possiate sentirvi rispondere è " Ci sento benissimo " . A parte il fatto che solo un perfetto imbecille potrebbe darvi una risposta del genere (che cosa c' entra il sentirci ? Voi state scrivendo , *mica* parlate) , esistono altri perfetti imbecilli che danno risposte del tipo [...]

[20c] Ma dicevo : ho riletto tutti gli scambi e non ho trovato nessuno , tra quelli che ti hanno risposto (che poi siamo stati in tre , *mica* una folla , e si fa presto a rileggere) che usasse i termini che hai usato tu fin da subito : " Saturno negativo " , " Marte sfigato " [...]

[20d] Possibile che ai comuni mortali si continuino a raccontare puttanate sul sudario di Cristo ? Oh , si raccontano puttanate un po' su tutto . Però le piramidi le hanno fatte davvero gli alieni , dai . Quello è chiaro , altrimenti perché la Piramide

di Cheope sarebbe altra esattamente un milionesimo la distanza che ci separa dal sole ? Esattamente , eh ! Sì sì , l' hanno misurata col metro flessibile (la distanza tra terra e sole , *mica* la piramide , per quella hanno dovuto usare metodi empirici) [...]

(b) relazioni di parafrasi {se uno non vuole può non leggere; *mica* è obbligato} [21a], {in giro; in strada} [21b], {pc muletto; *mica* chissà quale pc di ultima generazione} [21c], {non avere nulla a che vedere; partenogenesi; correlazione} [21d], {non desiderare figli, considerare ammissibile il divorzio, annullamento del vincolo civile; desiderare figli, credere nell'indissolubilità, contrarre un matrimonio civilmente valido} [21e], spesso mediate da un processo inferenziale [21f],

[21a] Chi non fosse d' accordo sui contenuti o sugli articoli stessi , bene può postarne altri come fa Calos e qualcun altro . Il bello del ng è proprio questo , che ognuno spara le sue cose . Certo ci si arrabbia , ci si scontra , ma è questo il sale della democrazia e della libertà . Se uno non vuole può non leggere (io per es. leggo si e no dei posti di Pierangelo , tanto me ne manda una copia in MP , gna faccio più :-))) , va boh , scherzavo) , *mica* è obbligato , può contraddire , può postare nuovi thread , può .... ok , anche " rinunciare " , deporre le armi

[21b] la conversazione si stava " avvitando " su se stessa ... Alternandosi tra il potenziale aiuto che avrei potuto dare agli altri e i miglioramenti che avrei potuto ottenere , come se il parlare di me si potesse tradurre automaticamente in aiuto al prossimo . Ho risposto che non mi interessava , nè condividevo un " aiuto " che si basa sostanzialmente su un fuoco di fila di domande personali fatte in giro da parte di estranei . Beh la risposta (disarmante) : " Estranei ? Ma noi ora ci conosciamo e comunichiamo ... e qui non siamo *mica* in strada ... "

[21c] Sfoglio il newsgroup con un Outlook eXpress con il pc muletto .. Usa Forte Agent che va una favola anche con un pc muletto (io ho un k-2 *mica* chissà quale pc di ultima generazione)

[21d] A mio parere il terrorismo islamico non ha nulla a che vedere con l' immigrazione . Ho già precisato che nel nostro paese ne abbiamo rilevato tracce e non manifestazioni . Ma queste tracce non si sono *mica* prodotte per partenogenesi ! Non c' è correlazione . Quando alle Olimpiadi di monaco è stata fatta la strage contro gli israelani , non c' era l' immigrazione di adesso , ma c' era il terrorismo . Quando Gheddafi ha combinato i casini con gli aerei non c' era l' immigrazione di adesso , ma c' era il terrorismo . Non c' è correlazione .

[21e] Perché se due vanno a dichiarare al tribunale ecclesiastico che il loro sacramento non era valido " perchè non desideravano figli " , o " perché consideravano ammissibile il divorzio " , questo deve consentire loro di annullare anche il vincolo civile ? La legge italiana non prevede *mica* che , per contrarre un matrimonio civilmente valido , sia obbligatorio desiderare figli o credere nell' indissolubilità !

[21f] hai idea se metteva le card clonate (e perfettamente funzionanti) sul mercato a l' una quanto ci guadagnava ? mi spiace che qualcuno si roda , ma quel tizio c' e' riuscito altrimenti non credo lo avrebbero arrestato (per l' arresto ci vogliono prove inconfutabili di reato *mica* si arresta cosi una persona solo perché aveva qualche card clonata per il condominio ...)

(c) relazioni, invece, in cui il contesto extra-linguistico e/o elementi del sapere enciclopedico sono necessari a precisare l'inferenza in gioco:

[22a] Io di Ratzinger non mi fiderei troppo (in vaticano viene chiamato con il soprannome del RATTO !) Ma quale busta è stata aperta ? Non è *mica* ex Rischiatutto : Signora Longari !

[22b] Inviato da non votate mai per Berlusca wrote : Ecco a voi Signore e Signori dagli Stati Uniti d' America il Presidente il Cavaliere il Dirigente SILVIO BERLUSCONI , è arrivato dopo un mese rifatto , un pò come M. Jackson eh sì il look prima e poi il LAVORO . Però la GIUSTIZIA mai prima del LAVORO . Certo non poteva farsi *mica* una legge che gli impedisse di rifarsi il look per poter essere il Presidente del Consiglio .

In riferimento alle più recenti tipologie di inferenze (vedi sopra), si ravvisano nei dati sia "forward inferences", provocate da un *trigger*, come *l'ascendente* nell'esempio [23],

[23] In passato , quando la religione era forte e la scienza debole , gli uomini confondono la magia per la medicina ; oggi , quando la scienza è forte e la religione è debole , gli uomini confondono la medicina con la magia (Thomas Szasz) . Il radiologo (riferendosi a una radiografia del colon) : " La signora l' ha fatto l' ascendente ? " . E la signora , tutta contenta : " Che bello , *mica* lo sapevo che in questo reparto facevate anche l' oroscopo ! "

sia "bridging inferences", tratte solo *a posteriori*, quando si rende necessario stabilire coerenza tra un segmento di discorso ed il discorso precedente:

[24] Per intenderci Tanzillo era quello che si chiedeva come un neonato potrebbe avere bisogno di una trasfusione , *mica* va in giro in moto !

3. CONCLUSIONI. La ricerca è ancora agli inizi, e lascia aperti molti quesiti (cfr. Hansen - Visconti *i.p.*). Due punti, tuttavia, emergono con chiarezza da questa prima ricognizione:

(j) lo studio di *mica* consente di giungere ad una tipologia fine delle possibili relazioni di un enunciato con il co-testo precedente, e quindi ad una più precisa caratterizzazione della dimensione testuale della datità;

(ij) *corpora* come i NUNC sono di estremo interesse per ricerche di tipo testuale, anche complesse.

BIBLIOGRAFIA.

BARBERA

2007 *i.s.* Manuel Barbera, *I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo*, in "Cuadernos de filología italiana" XIV (2007) 13-32, in corso di stampa.

¶ 1. Manuel Barbera, *Tra bmanuel.org e corpora.unito.it. Per la storia di un gruppo di ricerca*, in questo volume, pp. 3-20.

BENINCÀ et alii

1996 *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio Lepschy*, a cura di Paola Benincà, Guglielmo Cinque, Tullio De Mauro e Nigel Vincent, Roma, Bulzoni, 1996.

BERNINI - RAMAT

1992 Giuliano Bernini - Paolo Ramat, *La frase negativa nelle lingue d'Europa*, Bologna, il Mulino, 1992.

BIRNER

2006 Betty J. Birner, *Semantic and Pragmatic Contributions to Information Status*, in HANSEN - TURNER 2006, pp. 14-32.

BIRNER - WARD
i.s. *Drawing the Boundaries of Meaning: Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn*, edited by Betty J. Birner and Gregory Ward, Amsterdam, John Benjamins, in corso di stampa.

CHAFE
1976 Wallace Chafe, *Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View*, in LI 1976, pp. 25-55.

CINQUE
1976/91 Guglielmo Cinque, *Mica: note di sintassi e pragmatica*, in CINQUE 1991, pp. 311-323. Versione originale: G. C., *Mica*, in "Annali della facoltà di Lettere e filosofia dell'Università di Padova" I (1976) 101-112.
1991 Guglielmo Cinque, *Teoria linguistica e sintassi italiana*, Bologna, il Mulino, 1991.

COLE
1978 *Pragmatics*, edited by Peter Cole, New York, Academic Press, 1978. = "Syntax and Semantics" IX (1978).
1981 *Radical Pragmatics*, edited by Peter Cole, New York, Academic Press, 1981.

CORINO
¶ 13 Elisa Corino, *NUNC est disputandum. Aspetti della testualità e questioni metodologiche*, in questo volume, pp. 225-252.

DAHL
1979 Östen Dahl, *Typology of Sentence Negation*, in "Linguistics" XVII (1979) 79-106.

DRYER
1996 Matthew S. Dryer, *Focus, Pragmatic Presupposition, and Activated Propositions*, in "Journal of Pragmatics" XXVI (1996) 475-523.

FERRARI - DE CESARE
i.s. *Lessico, grammatica, testualità*, a cura di Angela Ferrari e Anna-Maria De Cesare, Basilea, Università di Basilea, 2007, in corso di stampa = "Acta Linguistica Basiliensa" XVIII.

GGIC I → RENZI - SALVI et alii 1988; II → RENZI - SALVI et alii 1991; III → RENZI - SALVI et alii 1995.

GIVÓN
1978 Talmy Givón, *Negation in Language: Pragmatics, Function, Ontology*, in COLE 1978, pp. 69-112.

HANSEN - TURNER
2006 *Explorations in the Semantic/Pragmatics Interface*, edited by Maj-Britt Mosegaard Hansen and Ken Turner, Copenhagen, Nordisk Sprog- og Kulturforlag, 2006 = "Acta Linguistica Hafniensa" XXXVIII (2006).

HANSEN - VISCONTI
i.s. Maj-Britt Mosegaard Hansen - Jacqueline Visconti, *On Reinforced Negation in French and Italian*, comunicazione al *Colloque "Perspectives contrastives et grammaticalisation". Université de Fribourg, Département de français, 2-3 octobre 2006*, Amsterdam, Elsevier "Studies in Pragmatics", in corso di stampa.
i.p. Maj-Britt Mosegaard Hansen - Jacqueline Visconti, *On the Diachrony of Reinforced Negation in French and Italian*, comunicazione al *[IPrA 10] 10th International Pragmatics Conference (Göteborg, 8-13 July 2007)*, in preparazione.

HORN
1989 Laurence R. Horn, *A Natural History of Negation*, Chicago, Chicago Un. Press, 1989.

JESPERSEN
1917 Otto Jespersen, *Negation in English and Other Languages,* Copenhagen, A. F. Høst, 1917.

LI
1976 *Subject and Topic. Papers from the Symposium on Subject and Topic Held in March, 1975 at the University of California, Santa Barbara*, a cura di Charles N. Li, New York, Academic Press, 1976.

LIZ 4.0 → STOPPELLI - PICCHI 2001

MANZOTTI - RIGAMONTI
1991 Emilio Manzotti - Alessandra Rigamonti, *La negazione*, in *GGIC* II, pp. 245-320.

MOLINELLI
1987 Piera Molinelli, *The Current Situation as Regards Discontinuous Negation in the Romance Languages*, in RAMAT 1987, pp. 165-172.

MUNITZ - UNGER
1974 *Semantics and Philosophy*, edited by Milton K. Munitz - Peter K. Unger, New York, New York University Press, 1974. [Seminar held in 1972-73 under the auspices of New York University Institute of Philosophy of the New York University, Dept. of Philosophy].

PARRY
1996 Mair Parry, *La negazione italo-romanza: variazione tipologica e variazione strutturale*, in BENINCÀ et alii 1996, pp. 225-257.

PRINCE
1981 Ellen F. Prince, *Toward a Taxonomy of Given-New Information*, in COLE 1981, pp. 223-255.
1992 Ellen F. Prince, *The ZPG Letter: Subjects, Definiteness, and Information-Status*, in THOMPSON - MANN 1992, pp. 295-325.

RAMAT
1987 *Linguistic Typology*, a cura di Paolo Ramat, Berlin, Mouton de Gruyter, 1987.

RENZI - SALVI et alii
1988 *Grande grammatica italiana di consultazione*. Volume I, *La frase. I sintagmi nominale e preposizionale*, a cura di Lorenzo Renzi, Bologna, il Mulino, 1988.
1991 *Grande grammatica italiana di consultazione*. Volume II, *I sintagmi verbale, aggettivale, avverbiale. La subordinazione*, a cura di Lorenzo Renzi e Giampaolo Salvi, Bologna, il Mulino, 1991.
1995 *Grande grammatica italiana di consultazione*. Volume III, *Tipi di frase, deissi, formazione delle parole*, a cura di Lorenzo Renzi, Giampaolo Salvi e Anna Cardinaletti, Bologna, il Mulino, 1995.

SCHWENTER
2003 Scott Schwenter, No *and* Tampoco*: a Pragmatic Distinction in Spanish Negation*, in «Journal of Pragmatics», 35, 7 (2003) 999-1030.
i.s. Scott Schwenter, *Fine-tuning Jespersen's Cycle*, in BIRNER - WARD *i.s.*

SERIANNI
1989/91 Luca Serianni, *Grammatica italiana. Italiano comune e lingua letteraria*, con la collaborazione di Alberto Castelvecchi, Torino UTET 1991_2 [1989_1] “Libreria”.

SIERWISKA - SONG
1998 *Case, Typology, and Grammar. In Honor of Barry J. Blake*, edited by Anna Sierwiska and Jae Jung Song, Amsterdam - Philadelphia, John Benjamins, 1998.

STALNAKER
1974 Robert Stalnaker, *Pragmatic Presuppositions*, in MUNITZ - UNGER 1974, pp. 197-213.

STOPPELLI - PICCHI
2001 *LIZ 4.0. Letteratura italiana Zanichelli. CD-ROM dei testi della letteratura italiana*, a cura di Pasquale Stoppelli ed Eugenio Picchi, Bologna, Zanichelli, 2001, quarta edizione per Windows.

THOMPSON
1998 Sandra A[nnear] Thompson, *A Discourse Explanation for the Cross-Linguistic Differences in the Grammar of Incorporation and Negation*, in SIERWISKA - SONG 1998, pp. 307-40.

THOMPSON - MANN
1992 *Description: Diverse Analyses of a Fundraising Text,* edited by Sandra A[nnear] Thompson and William Mann, Amsterdam, John Benjamins, 1992.

VISCONTI
i.s. Jacqueline Visconti, *Lessico e contesto: sulla diacronia di* mica, in FERRARI - DE CESARE *i.s.*

ZANUTTINI
1997 Raffaella Zanuttini, *Negation and Clausal Structure: A Comparative Study of Romance Languages*, Oxford, Oxford University Press, 1997 "Oxford Studies in Comparative Syntax".

CORPORA E SITI DI RIFERIMENTO.

NUNC `http://www.bmanuel.org/projects/ng-HOME.html`

OVI db testuale `http://ovisun198.ovi.cnr.it/italnet/OVI/`

MARCO CARMELLO
(Università di Torino)

21. "Dovere" deontico e "dovere" anankastico fra semantica e pragmatica. *Una ricerca* corpus-based.

0. PREMESSA. Nel quadro di una logica deontica, ossia di una logica che discuta la natura della norma e del *dovere* – come quella proposta, ad esempio, da von Wright –, la differenza anankastico / deontico è non solo chiara, ma anche fondamentale. Per lo scopo del presente scritto si potrebbe rappresentare tale differenza come segue[1]:

(a) il deontico predica un dovere normativo, ossia pone una richiesta che deve essere rispettata, in maniera vincolante, se non si vuole commettere un'infrazione (senso del greco *τὸ δέον*);

(b) l'anankastico predica una necessità normativa, ossia pone una condizione che deve verificarsi nella realtà perché l'atto sia valido (senso del greco *ἀνάγκη (ἐστί)*)[2].

Vi è quindi un'evidente differenza semantica fra anankastico e deontico[3], che, come hanno suggerito M.-E. Conte e A. G. Conte, deve esser tenuta presente anche nell'analisi delle lingue naturali, distinguendo i contesti in cui alcuni predicati, come "dovere" od "essere necessario", sono anankastici da quelli in cui sono deontici.

Linguisticamente però le cose si complicano, almeno riguardo "dovere", predicato su cui si concentra la ricerca. I dati sembrano infatti suggerire che l'uso non epistemico di "dovere" crei un contesto logico di necessità, da interpretarsi, nel calcolo del primo ordine, come $\Box P(x)$ (o **N** $P(x)$; ossia necessariamente P di x). Semanticamente questa sembra essere l'interpretazione adeguata sia per gli enunciati anankastici come: "i candidati non devono avere più di diciotto anni", sia per quelli deontici come: "i candidati non devono copiare".

Chiaramente questa situazione non può soddisfare la nostra intuizione di parlanti italiani: tutti noi infatti comprendiamo che mentre il primo enunciato predica un requisito, una necessità normativa, il secondo invece predica un preciso dovere. Eppure in italiano la differenza deontico / anankastico viene oscurata dal predicato "dovere", che "fonde" le due modalità su di un'unica, indifferenziata, idea di necessità.

La differenza logica fra anankastico e deontico non sembra dunque appartenere al sistema semantico della lingua italiana, ma pare piuttosto muoversi a livello pragmatico[4]; infatti solo se

---

[1] Tenga presente il lettore che l'immagine della differenza deontico/anankastico da me tracciata è quella, sperabilmente, utile ad un linguista, il logico deontico potrebbe quindi non condividerla pienamente.

[2] Un esempio, di cui sono debitore ad Amedeo Giovanni Conte può contribuire a chiarire meglio il concetto. In Arabia Saudita le donne non devono guidare e non devono votare; mentre però il primo "dovere" è deontico, ossia se una donna guida commette un'infrazione perseguita dalla legge, il secondo è anankastico: una donna può quindi, senza commettere alcuna infrazione, anche votare, solo il suo voto non vale perché le manca, secondo la legge araba, il prerequisito necessario per esprimere il voto, ossia l'essere uomo.

[3] Come del resto suggerisce anche una considerazione etimologica: l'*ἀνάγκη* è la costrizione ineluttabile, il destino cui l'uomo non può fare a meno di obbedire, quindi il dovere che esprime non è sanzionante ma necessitante: o si fa così o si è addirittura esclusi dall'applicazione della norma (il significato greco suggerisce che si fa così perché non si potrebbe fare altrimenti). Il *δέον* suggerisce invece l'idea di appropriatezza, convenienza, idoneità, è quindi un dovere sanzionante ma non necessitante: si può anche non fare così, ma si può essere puniti.

[4] Non ho qui spazio sufficiente per meglio definire quanto dico, mi pare però doveroso aggiungere in questa nota alcune parole sulla natura della differenza anankastico / deontico. Che questa differenza abbia un carattere logico mi pare evidente, anche se implicitamente, da quanto detto sopra; dico che però essa è, in italiano, pragmatica e non semantica perché in italiano non esistono due predicati distintamente dedicati al dovere anankastico ed al

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 347-362.

abbiamo un'informazione pragmatica sufficientemente ricca possiamo individuare l'anankasticità o la deonticità del predicato: definisco quindi come logico-pragmatica la natura della differenza anankastico / deontico in italiano.

0.1 *Dovere, Potere, Vietare*: per un possibile test di parafrasi. L'affermazione della non semanticità della differenza anankastico/deontico va moderata: sembrano infatti esserci alcuni test linguistici capaci di distinguere le occorrenze anankastiche e quelle deontiche di "dovere".

Uno di questi test è quello della ripresa anaforica proposto da Maria-Elisabeth Conte (Conte, M.-E., 1993, pp. 5-9), per il quale rimando alle pagine della Conte stessa.

Qui cercherò di presentare un altro possibile test di interpretazione – d'ora in poi chiamerò così quei test che permettono di selezionare l'anankasticità o la deonticità di "dovere" –: il "test di parafrasi". Quanto dirò a proposito di questo possibile test è ancora un abbozzo: non si tratta quindi di una vera e propria proposta, ma di un'idea sottoposta al giudizio, oltre che alla pazienza, del lettore[5].

Partiamo dal seguente enunciato,

[1] È vietato fumare.

che ha un'interpretazione univocamente deontica, esso infatti predica un dovere: se parafrasiamo [1] con *dovere*, otteniamo l'enunciato [2],

[2] Non si deve fumare

che ha sì un'interpretazione deontica, ma più sfumata, come dimostra la possibilità di poter usare [2] ma non [1] in contesti che hanno valore anankastico. Si confronti ad esempio la legittimità di un enunciato come "I candidati alla presidenza della lega anti-fumo devono avere almeno diciotto anni e non devono fumare", con l'illegittimità di "I candidati alla presidenza della lega anti-fumo devono avere almeno diciotto anni ed è loro vietato fumare".

Nel primo caso si descrivono le caratteristiche necessarie per concorrere al posto di presidente della lega anti-fumo, cosa non solo legittima ma anche necessaria in qualsiasi bando di concorso.

Nel secondo caso alla definizione di una caratteristica viene accompagnata la predicazione di un dovere, cosa che potrebbe non essere considerata pienamente legittima.

In cosa consiste dunque la "maggior forza" deontica di "vietare"? La risposta sembra essere nella semantica del predicato "vietare", il quale – come mi sembra – non solo non predica mai l'impossibilità di fare qualcosa – come talora fa invece "dovere" –, ma addirittura predica sempre e solo il divieto di fare qualcosa che sarebbe possibile fare, come dimostrano [3] e [4]:

[3] Qui si potrebbe fumare, ma è vietato

[4] Una volta nei cinema si poteva fumare, ma oggi è vietato

---

dovere deontico, né vi sono fenomeni sintattici tali da determinare una differenza strutturale fra le due interpretazioni. Le lingue, in generale, possono marcare semanticamente la differenza anankastico/deontico – ho, ad esempio, il sospetto che la differenza, esistente in greco classico, fra *ἀνάγκη (ἐστί)* e *δέον (ἐστί)* sia di questo tipo –, ma, come nel caso di altre differenze semantiche, possono anche non farlo. Ora credo che una differenza logica non semanticamente marcata possa essere recuperata dalla lingua in altri àmbiti: ed in italiano la differenza anankastico/deontico è recuperata in àmbito pragmatico.

[5] Una franca ed utile discussione avuta con Amedeo G. Conte mi ha rinsaldato nelle opinioni che esprimerò. A Conte devo dunque non solo il ringraziamento per l'attenzione, ma anche quello per l'invito alla profondità ed al rigore della riflessione. Sappia il lettore che il molto di inesatto, forse di errato, che ancora resta in quanto scrivo è però mia esclusiva responsabilità.

In entrambi gli enunciati possiamo sostituire a "è vietato" "non si deve", in [4] possiamo addirittura usare "potere". Traduco quindi le avversative degli esempi con: 'ma non si deve' (es. [3]); '…ma oggi non si deve / non si può (più)' (es. [4]). Nelle due parafrasi "dovere" indica qualcosa che è possibile **e** vietato fare: si noti che non esiste una causa per così dire "naturale" che impedisca di fumare laddove è vietato farlo, come tristemente dimostra la frequente infrazione del divieto di fumo.

Ci si deve però chiedere perché in [4] "vietare" possa essere sostituito anche con "potere". Anzitutto credo che la possibilità sia data dalla differenza di riferimento temporale fra le due frasi (una volta…ora), poi si deve considerare che anche per "potere" sembra verificarsi una vischiosità semantica in qualche modo analoga[6] a quella notata nell'introduzione (cfr. § 0) per "dovere".

Anche "potere" nel calcolo del primo ordine è reso semplicemente con: $\Diamond P(x)$ (o **P** P(x); ossia 'possibilmente P di x'), con problemi simili a quelli di "dovere" per quanto riguarda anankastico e deontico. La semantica della possibilità permette però una lettura in termini di liceità, da cui segue la possibilità di parafrasare le due occorrenze di "potere" negli esempi [3] e [4] e quella proposta come alternativa di "vietare" in [4] con "è lecito", la qual cosa penso risolva alcuni dei nostri problemi[7].

Possiamo valutare ancor meglio la differenza fra "vietare", "dovere" e "potere" considerando i seguenti tre enunciati:

[5a] Quando si è completamente immersi in acqua è vietato fumare
[5b] Quando si è completamente immersi in acqua non si deve fumare
[5c] Quando si è completamente immersi in acqua non si può fumare

[5a] e [5c] non sono problematici: [5a] è evidentemente assurdo, a riprova dell'impossibilità di vietare qualcosa che non si può fare, [5c] invece è una descrizione adeguata della realtà fisica del mondo; e [5b]? L'enunciato [5b] è problematico perché può essere una buona parafrasi sia di [5a] che di [5c]: se la predicazione di necessità di "dovere" verte sul divieto e non sulla possibilità allora, come si è fin qui visto, "dovere" parafrasa "vietare". Tuttavia la predicazione di necessità può vertere anche sul "reale", ed in questo caso "dovere" parafrasa "potere" di [5c]: se è possibile od impossibile fare qualcosa, allora è anche doveroso farla o non farla[8].

Nonostante quanto detto, [5b] come parafrasi di [5c] suona forzato: [5c] infatti parla della realtà fisica del mondo. Se però usassimo enunciati esprimenti regole o realtà differenti da quella di cui si parla nella batteria di esempi [5a-c] la parafrasi "potere"/"dovere" sarebbe regolare[9]; si considerino:

[6a] Se giochi a scacchi puoi muovere l'alfiere in diagonale
[6b] Se giochi a scacchi devi muovere l'alfiere in diagonale

---

[6] Insisto con forza sull'aggettivo: "analogo" non significa uguale e neanche semplicemente simile, ma 'funzionante secondo regole reciprocamente traducibili'. Possiamo quindi usare "potere" come chiave di lettura di "dovere" e viceversa, ma dobbiamo sapere che il comportamento dei due predicati differisce secondo una proporzione determinabile.

[7] Il discorso dovrebbe essere più articolato, ma sono costretto a rimandare ad un'altra sede.

[8] Mario Squartini ha richiamato la mia attenzione sulla possibilità di introdurre la modalità aletica come utile categoria interpretativa per gli esempi [5a-c]. Ritengo l'osservazione interessante perché, come avrò di notare alla nota 9, l'anankastico sembra avere un ruolo importante nel passaggio dall'uso epistemico a quello deontico del predicato "dovere"; tuttavia devo rimandare un'analisi più ricca di questo tema ad un'altra sede.

[9] Si pone qui un problema affascinante e difficile: quali sono i rapporti specifici fra modalità anankastica ed uso epistemico di "dovere"? Sarebbe forse possibile dire che l'anankastico funziona come un ponte fra il deontico e l'epistemico? Perché, nel caso dell'anankastico, dobbiamo parlare di regole o di realtà latamente "normative" e non, come ho fatto, in [5a], [5b] e [5c] di realtà fisiche *et similia*? Chiaramente non posso far altro che rilevare il problema, e rimandare ad altra sede la discussione.

[7a] Per contattarci gli ascoltatori possono chiamare il seguente numero telefonico
[7b] Per contattarci gli ascoltatori devono chiamare il seguente numero telefonico[10]

Nel caso di [6] la possibilità di muovere l'alfiere in diagonale implica, nel gioco degli scacchi, una necessità, o, altrimenti detto, è impossibile, giocando a scacchi, muovere l'alfiere in altro modo. Si badi che non è un'impossibilità assoluta: evidentemente si potrebbe pensare di muovere l'alfiere in verticale, od in orizzontale, od a Γ, solo che se muovessimo l'alfiere in questa maniera non giocheremmo più a scacchi, ma staremmo praticando un gioco diverso che si serve della stessa scacchiera e degli stessi pezzi usati dagli scacchi. [6] sta insomma definendo quelle che A. G. Conte chiamerebbe le "regole anankastico-costitutive" del gioco degli scacchi: "dovere" dunque parafrasa "potere" predicando che è necessariamente possibile che l'alfiere si muova in diagonale ($\Box\Diamond P(x)$), perché gli scacchi sono quel gioco definito da un determinato gruppo di regole, una delle quali stabilisce che l'alfiere muova in diagonale e non altrimenti.

Anche in [7], sebbene non si parli – almeno non direttamente[11] – di regole anankastico-costitutive, il rapporto fra "dovere" e "potere" funziona come in [6]: per poter partecipare al programma gli ascoltatori devono chiamare, quindi la possibilità è necessitata. Nel caso del programma radiofonico possiamo essere solo uditori, come nel caso di una partita a scacchi possiamo essere solo spettatori, ma se vogliamo partecipare al programma dobbiamo dare seguito alla possibilità di telefonare.

Quello che si è venuti dicendo potrebbe già bastare, se non fosse che la scelta di usare "vietare" come chiave di interpretazione per l'uso deontico di "dovere" potrebbe indurre in errore, facendo interpretare il deontico come una mera sanzione ('o fai così o sei punito').

Che non sia così è dimostrato da [8a], che è un imperativo (categorico) la cui verità non è negata né da [8b] né da [8c]:

[8a] Devi pensare prima di parlare
[8b] È possibile pensare prima di parlare
[8c] È possibile non pensare prima di parlare

Sia [8a], sia [8b], sia [8c] possono essere simultaneamente veri: è vero che si predica un dovere obbligante (deontico) consistente nel pensare prima di parlare; è vero che, come dimostrano pochi saggi, è possibile ottemperare a questo dovere; è vero, come dimostrano più di 100.000 anni di storia umana, che è possibile non farlo.

La situazione di [8] dipinge quindi un uso di "dovere" parallelo a quello tratteggiato per l'uso di dovere deontico parafrasabile con "vietare": come in quel caso era impossibile vietare ciò che è già impossibile, in questo è impossibile prescrivere ciò a cui è impossibile ottemperare (si pensi alla stolidità di un ordine come: "sott'acqua devi respirare a pieni polmoni").[12]

La differenza fra interpretazione anankastica e deontica sembra dunque consistere nella differenza di rapporto fra *possibilità* e *necessità*, per cui la possibile forma logica del deontico sembrerebbe essere [9a], mentre quella dell'anankastico pare [9b]:

[9a] $(\Diamond P(x) \wedge \Diamond\neg P(x)) \wedge \Box\, (P(x) \vee \neg P(x))$
[9b] $\Box\, (\Diamond P(x) \vee \Diamond\, \neg P(x))$

[10] L'esempio [6] ha un'evidente ascendenza wittgensteiniana, sull'esempio [7] ha invece richiamato la mia attenzione Amedeo G. Conte.

[11] In realtà ritorna qui il problema del rapporto anankastico-epistemico: si potrebbe dire che la possibilità di fare chiamate telefoniche faccia parte delle caratteristiche proprie di alcuni programmi radiofonici – chiamiamoli programmi a microfono aperto –; si potrebbe quindi dire che un programma a microfono aperto sia definibile come tale solo se gli ascoltatori possono chiamare; si potrebbe dunque evincere che la possibilità di chiamata sia una regola anankastico-costitutiva del programma a microfono aperto; si potrebbe così concludere che se gli ascoltatori vogliono partecipare al gioco "programma radiofonico a microfono aperto" devono chiamare.

[12] Se dunque il "devi" di [8a] fosse anankastico, il mondo sarebbe ancora un Eden, ossia un paradiso.

Le due formule dicono quanto segue:

[9a'] 'sono possibili tanto P(x) quanto la sua negazione (¬P(x)), ma è necessario solo o P(x) o la sua negazione' (puoi fumare o non fumare, ma devi non fumare; puoi pensare o non pensare, ma devi pensare);

[9b'] 'è necessario che sia possibile solo o P(x) o la sua negazione' (necessariamente è possibile muovere l'alfiere in diagonale, ed è possibile solo questo tipo di mossa dell'alfiere se si vuole giocare a scacchi).

Questa è dunque la proposta di test di parafrasi che propongo al lettore: come si nota la cosa è ancora abbozzata, penso però sia utile avere a disposizione un *discrimen* che permetta, almeno in linea di principio, una classificazione non impressionistica di anankastico e deontico.

Resta da dire perché ritenga che questa differenza logica sia pragmaticamente veicolata e non appartenga all'àmbito della semantica di "dovere".

Anche in questo caso mi limito solo ad alcuni accenni: come il lettore avrà intuito, più di quanto io abbia effettivamente dimostrato, l'opposizione anankastico / deontico può essere sostanzialmente riportata ad una questione di àmbito ("*scope*") del funtore di necessità. Nel caso dell'anankastico il funtore di necessità lega quello di possibilità, nel caso del deontico no: nell'anankastico dunque il funtore di possibilità è legato e solo quello di necessità è libero.

La semantica di "dovere" però ci dice solamente che il predicato italiano *dovere* svolge il ruolo di funtore di necessità, senza dirci nulla riguardo al suo *scope*; detto in termini più banali, quando usiamo "dovere" sappiamo preventivamente che qualcosa verrà necessitato[13], ma cosa esattamente sarà necessitato ci verrà detto solo dal contesto informativo.

È dunque la pragmatica a darci, anche se non sempre, le informazioni necessarie a determinare l'àmbito applicativo del funtore di necessità *dovere*. In questo senso dico che la differenza anankastico/deontico appartiene al livello pragmatico e non quello semantico della lingua.

1. L'USO DEI CORPORA.

La particolare natura logico-semantica della distinzione deontico / anankastico rende auspicabile una ricerca *corpus-based*, vuoi perché tale ricerca permette, se obbedisce ai requisiti che elencherò, di ricostruire i contesti di proferimento di "dovere", vuoi perché essa consente al ricercatore di controllare diverse tipologie testuali, differenti registri linguistici, quindi, in sintesi, diversi settori di lingua. Ecco perché corpora e non corpus: ritengo infatti utile determinare come l'opposizione anankastico/deontico viva nella lingua in genere, anche al di fuori di quegli àmbiti normativi che ne rappresentano il campo privilegiato.

Perché però la ricerca sia valida i corpora usati devono avere alcuni imprescindibili requisiti.

1.1 REQUISITI DEI CORPORA.

I requisiti che poniamo sono sostanzialmente tre.

(**a**) Il corpus deve permettere di avere output[14] di estensione rilevante. Essendo essenziale il contesto di proferimento per determinare se un'occorrenza di "dovere" sia anankastica o deontica, è necessario che gli output della ricerca abbiano un numero di parole minimo capace di garantire la presenza di tutte le informazioni necessarie. Per definire il contesto minimo, in realtà, sarebbero pertinenti il numero di unità testuali (o di *speech acts*) e non il numero di parole, ma,

---

[13] Od almeno sappiamo che così è per gli usi non epistemici di "dovere" (sebbene sia propenso a pensare che così, in realtà, sia anche per gli usi epistemici).

[14] Scelgo volutamente un termine neutro come output per due buoni motivi: il primo è l'uso giustamente formale del concetto di token fatto da Barbera in questo volume (cfr. *supra* Barbera - Corino - Onesti ¶ 3, § 1.3). Il secondo motivo invece potrebbe essere così formulabile: la possibilità di ampliare la citazione per poter meglio definire la natura di un'occorrenza di "dovere" comporta anche l'esistenza di due token diversi, od il token è sempre uno? A mio dire, se si segue il discorso di Barbera e si valutano correttamente le citazioni che egli trae da *Quiddities* di Quine ci sono buone ragioni per rispondere che il token rimane sempre uno. Introdurre qui il concetto di token richiederebbe quindi un ulteriore appesantimento teorico da parte mia, per ciò rinuncio e, per tutte le questioni inerenti a questo concetto, rimando al già citato articolo, ¶ 3, § 1.3.

con comoda approssimazione statistica, si potrebbe dire che perché un output sia rilevante dovrebbe essere costituito da almeno una settantina di parole.

Perché questo requisito sia soddisfatto in maniera ottimale è bene che i corpora usati nella ricerca permettano un eventuale ampliamento dell'output: possono esservi infatti casi dubbi ma risolvibili in presenza di contesti che eccedano le settanta parole stabilite. A questo riguardo tuttavia val la pena specificare che un contesto di settanta parole – talora anche meno – sembra in genere essere sufficiente per determinare la natura di un occorrenza di "dovere"; se infatti non si riesce a determinare il valore anankastico o deontico di "dovere" con settanta parole, possiamo perlopiù dire di essere in presenza di un'occorrenza indecidibile.

Potremmo chiamare questa "legge dell'occorrenza minima e massima": al di sotto delle settanta parole è raro che occorrano tutte le informazioni pragmatiche utili alla determinazione del predicato, al di sopra la ricchezza di contesto non aggiunge di solito nulla a tali informazioni[15].

Un'ampiezza di settanta parole non inficia teoricamente la rilevanza dell'uso del corpus: si tratta di una finestra citazionale ancora contenuta per la maggior parte delle tipologie testuali.

(**b**) Un'altra condizione indispensabile è che i corpora offrano un numero di esempi sufficientemente ricco. Se gli esempi fossero, nel complesso, meno di un centinaio rischieremmo di non veder rappresentate alcune tipologie testuali possibili.

(**c**) La necessità di considerare differenti tipologie testuali mi porta a porre un'ultima condizione riguardante, a rigore, non il corpus ma il suo uso: la ricerca deve essere compiuta su corpora composti a partire da differenti tipologie di testo, alcune caratteristiche del testo possono infatti avere importanti conseguenze sul contesto di proferimento di "dovere".

Poiché è il sistema pragmatico della lingua a veicolare la differenza anankastico / deontico, è necessario, considerando la varianza dell'informazione pragmatica a seconda del contesto, che i corpora presi in esame, pur se costituiti da testi di diversa ed eterogenea differenza[16], rappresentino comunque le opposizioni normativo / non normativo e formale / informale[17].

1.2 NORMATIVO/NON NORMATIVO E FORMALE/INFORMALE. Sinora ho genericamente parlato di "informazione pragmatica", è però venuto il momento di dire che con questa espressione intendo quel genere di conoscenza che gli interlocutori possono ricavare dalla conversazione medesima applicando le massime conversazionali di Grice. L'applicazione di queste massime (quantità, qualità, relazione, modalità: cfr. Grice 1989/93) generalmente obbedisce ad un criterio di economia – il griceano principio di collaborazione – che potrebbe essere sintetizzato dal seguente *slogan*: "Dì, nel modo più adeguato possibile, tutto quello che serve a rendere limpida ed onesta la comunicazione, nulla di più, nulla di meno"[18].

Questo *quantum* necessario però varia a seconda del tipo di comunicazione, di interlocutore, di testo: ciò che sarebbe considerato prolissità in un testo comune non lo è in uno normativo, quel che parrebbe affettazione in una conversazione informale, diviene adeguato in una comunicazione formale. In poche parole, i testi formali e quelli normativi sono, anche se in maniera differente l'uno dall'altro, più dispendiosi, più ricchi di informazione pragmatica, rispetto quelli non normativi ed informali.

Considerando quanto detto sulla natura pragmatica della distinzione logica anankastico / deontico, è quindi chiaro perché si debbano avere corpora capaci di ben rappresentare le opposizioni formale / informale e normativo / non normativo.

---

[15] Nella presente sede il lettore dovrà accontentarsi di questa secca enunciazione: in realtà la "legge" dipende da quanto dicevo sopra (fine § 0.1) riguardo la determinazione d'àmbito del funtore di necessità.

[16] Chiaramente alcuni tipi di testo – quelli giuridici sono l'esempio migliore – si presteranno più di altri al nostro tipo di ricerca.

[17] La classica opposizione scritto / parlato ha in questa sede meno importanza.

[18] Per quanto qui detto rimando a Grice (Grice 1989/93, capitolo secondo). Il lettore noterà che nella mia semplicistica riscrittura del principio di collaborazione dò la prevalenza alle massime di quantità e di relazione.

1.2.1 DEFINIZIONE DELLE OPPOSIZIONI. Per normativo si intende un testo che esprime una serie di norme o regole. In questo senso sono normativi tanto il testo di una legge, quanto una ricetta di cucina, anche quest'ultima infatti offre istruzioni che devono essere seguite. Anche le istruzioni, le indicazioni di comportamento date a voce sono da considerarsi testi normativi. Sebbene a rigore siano testi metanormativi, ai fini di questa ricerca devono essere considerati normativi anche i testi che discutono quelli strettamente normativi, ad es.: le glosse di commento ai testi giuridici, le motivazioni di sentenza, o, passando all'orale, le arringhe degli avvocati, gli interventi parlamentari ecc. Altri esempi di testi normativi possono essere considerati le istruzioni tecniche, gli articoli sul *bon ton*, le prescrizioni mediche, in genere appunto ogni testo che contenga, norme, regole, istruzioni.

La definizione di formale è invece più vaga[19]. Per quanto riguarda la distinzione anankastico / deontico è possibile dare di "formale" una definizione classica: sono formali quei testi che rispettano appieno le regole e le convenzioni grammaticali dell'italiano standard e sono relativamente poco influenzati da usi gergali.

1.3 TIPOLOGIE TESTUALI UTILI. L'interferenza delle opposizioni formale / informale, normativo / non normativo, ci porta ad individuare quattro differenti tipologie testuali (sei se si vuole considerare anche l'opposizione scritto / parlato, che qui però è un sottotipo) capaci di esaurire il campo dell'opposizione anankastico/deontico.

(a) **Normativo-formale**: si tratta di testi come leggi, decreti, commenti giuridici, trattati di morale, codici deontologici, interventi parlamentari (orale), arringhe (orale), prediche (orale).

(b) **Normativo-informale**: possono essere ricette di cucina, rubriche di consigli, galatei di Donna Letizia, rimproveri (orale), istruzioni di esecuzione (sia orale sia scritto).

(c) **Non-normativo-formale**: pubblicazioni ufficiali, annuari, riviste di ricerca scientifica, prosa d'arte, lezioni (orale), commemorazioni (orale), prolusioni (orale), discorsi pubblici (orale) ecc.

(d) **Non-normativo-informale**: lettere private, newsgroup, scritti informali, qualsiasi testo orale che non rientri nelle categorie su riportate.

2. LA RICERCA. Per la mia ricerca ho usato i corpora approntati dall'*équipe* coordinata da Carla Marello e Manuel Barbera presso bmanuel.org e l'Università di Torino; sono questi infatti gli unici corpora, di cui io sia a conoscenza, in grado di rispettare tutte le caratteristiche richieste al § 1[20]. I corpora di cui mi sono maggiormente servito sono i seguenti:

(a) Jus Jurium (JUS), non ancora presente su internet: comprende codici, leggi statali e regionali, sentenze, trascrizioni stenografiche di interventi parlamentari, ecc.; risponde alla tipologia normativo-formale.

(b) Athenaeum (A): comprende testi tratti dalla rivista ufficiale dell'Università degli Studi di Torino; risponde alla tipologia non normativo-formale.

(c) NUNC cucina (NC): contiene testi tratti da un newsgroup di cucina, in cui abbondano ricette e consigli per la preparazione, conservazione ed il giusto uso di cibi e bevande; risponde alla tipologia normativo-informale.

(d) NUNC motori (NM); contiene testi tratti da newsgroup in cui si parla di motociclismo ed automobilismo; risponde alla tipologia non normativo-informale.

(e) Sono stati usati anche: NUNC fotografia (NF), che equivale sostanzialmente a NUNC motori, per arricchire la raccolta dati, e NUNC generico (NG).

[19] Tale definizione può poi diventare ardua quando si abbia a che fare con corpora basati su blog o newsgroup.

[20] Un grande vantaggio, nell'uso di questi corpora, consiste nella possibilità di poter ampliare la lunghezza della citazione a piacere, col risultato di poter valutare attentamente anche quei risultati che, a prima vista, potrebbero parere indecidibili.

L'uso di NUNC generico è importante, perché in esso sono raccolti testi provenienti da diversi newsgroup, alcuni dei quali formali, altri no, molti non normativi, ma alcuni normativi. NUNC generico funziona quindi come un corpus di controllo per i dati tratti.

La discussione degli esempi sarà suddivisa in tre sezioni: Deontici puri, Anankastici puri, Contesti incerti e cooccorrenze. Nella discussione il numero di esempi tratto da JUS, ancora incompleto, è limitato ad uno (più uno in nota): all'analisi più compiuta di questo materiale si riserverà un'altra sede[21].

2.1 DEONTICI PURI. Consideriamo gli otto esempi di sèguito:

[10a] [...] similmente decretiamo e dichiariamo che le presenti lettere in nessun tempo **potranno** venir revocate o diminuite , ma stabili sempre e valide **dovranno** perseverare nel loro vigore [...] NG,

[10b] [...] il responsabile dello stabilimento [...] **dovrà** conservare i registri per almeno tre anni [...] NA,

[10c] [...] niente caffè e distillato perché penso che **si debba** concludere con il vino dolce [...] NC,

[10d] [...] ora **devo** andare , mi sta chiamando, e quando lei chiama io corro [...] NG,

[10e] [...] nei messaggi inviati al newsgroup l' oggetto **dovrebbe** prima descrivere la marca ed il modello [...] NM,

[10f] [...] se vi fosse all' opposto una concezione economica seria, **dovrebbe** essere lo Stato a stampare le proprie banconote [...] NG,

[10g] [...] senza l' impiego delle mani che **dovranno** essere saldamente ancorate al manubrio [...] NM,

[10h] [...] non si può lasciare la scelta la caso , l' autore **DEVE** fare le sue scelte [...] NF.

Possiamo raggruppare i deontici ricorrendo alla classica tripartizione kantiana: categorico, pragmatico, ipotetico. Brevemente definisco come segue i tre tipi di deontico (cfr. Conte, A. G. 1999): categorico è quel deontico che predica un obbligo assoluto; pragmatico è quel deontico che predica un obbligo legato ad una data funzione del soggetto; ipotetico è quel deontico che predica un dovere che si dà se si verifica una data situazione.

Secondo questa tripartizione [10a] e [10g] sono deontici categorici, predicano infatti una disposizione assoluta, che viene intesa come atemporale ed universalmente valida[22]. [10b] è il classico deontico pragmatico: in questo caso il dovere di conservare i registri è inerente alla

[21] La scelta di non riportare altri esempi tratti da Jus Jurium è determinata da scrupulo filologico, ma ha una conseguenza notevole: dal nostro orizzonte viene quasi totalmente eliminato il registro normativo-formale. Poichè è al registro normativo-formale, particolarmente al suo più forte rappresentante, il testo giuridico, che solitamente è indirizzata la ricerca sulle modalità anankastica e deontica, l'effetto di tale elisione è quello di spostare la ricerca su linguaggi non specialistici, dimostrando così la vitalità di queste modalità anche al di fuori dei contesti che si penserebbero loro esclusivi.

[22] L'esempio [10a] è tratto da NG ed è una traduzione, abbastanza fedele, di un passo della bolla papale *Quo primum tempore*, che Pio V fece precedere nel 1570 al messale stabilito in base ai *decreta Tridentina* (*Missale Romanum ex decreto Sacrosanti Concilii Tridentini restitutum*). Il testo latino è il seguente: «praesentesve litterae ullo unquam tempore revocari, aut moderari possint, sed firmae semper et validae in suo existant robore, similiter statuimus, ac declaramus» (Pio V 1570/1904). Si noti però un fatto importante: il traduttore italiano rende con l'ausiliare "dovere" un congiuntivo iussivo del latino (Ernout - Thomas 1953 definiscono questo tipo di congiuntivo *subjonctif de volition* (p. 231), e dicono, riguardo alla terza persona, che tale congiuntivo «a échu l'expression de l'ordre» (p. 234), ma curiosamente sceglie il tempo futuro, sia per rendere *existant*, sia per tradurre *possint*. Riprenderò quest'esempio più avanti.

funzione di direttore di uno stabilimento. [10f] è un deontico ipotetico tipico: data una certa concezione economica, ne segue un preciso dovere.

Gli esempi rimanenti svelano come non sia sempre facile individuare con quale tipo di deontico si abbia a che fare. [10c] e [10e] paiono essere deontici ipotetici: in entrambi i casi lo scrivente enuncia quello che lui, in base a sue convinzioni, ritiene essere un obbligo, si può quindi concludere che l'obbligo dipenda dal verificarsi nella realtà di condizione che sono, al momento di enunciazione, solo *in mente*. Se così fosse ci troveremmo di fronte ad un vero e proprio deontico ipotetico. Credo che anche [10h] possa essere considerato un deontico ipotetico: il locutore infatti subordina l'obbligo ad un suo implicito pensiero che funge da precondizione dell'obbligo stesso[23].

[10d] all'apparenza parrebbe funzionare come un deontico categorico, ma a differenza di questo non predica un obbligo assoluto; l'assolutezza dell'obbligo infatti dipende anche dalla sua universalità, mentre nell'esempio il locutore predica un obbligo esclusivamente personale.

Tutti gli esempi fin qui visti obbediscono alle condizioni fissate in § 0.1 per il deontico: il responsabile di stabilimento potrebbe non conservare i registri, le mani potrebbero non essere saldamente ancorate al manubrio, le banconote potrebbero non essere direttamente stampate dallo Stato – come accade in realtà –, e così via per tutti gli altri esempi.

Resta forse ambiguo solo l'esempio [10a]: se infatti è vero che le lettere in questione potrebbero non rimanere sempre stabili e valide – come dimostra l'esigenza di porre l'obbligo che tali rimangano –, è però altrettanto vero che la bolla predica prima quella che pare essere un'impossibilità di tipo anankastico, per cui il "dovere" potrebbe essere visto anche come un non poter fare altrimenti. Riprenderò la discussione nel § 2.3.

2.2 ANANKASTICI PURI. Si considerino anche in questo caso alcuni esempi

[11a] `[...] le strade di cui al comma 2` **`devono`** `avere le seguenti caratteristiche minime [...]` JUS,

[11b] `[...] ogni composizione, per essere buona,` **`deve`** `avere tre requisiti [...]` A,

[11c] `[...] fatta salva la possibilità di realizzare entro Aprile 2004 all'interno dell'azienda dei locali riservati ai fumatori dei locali che` **`devono`** `avere le seguenti caratteristiche [...]` NC,

[11d] `Inutile sottolineare che Stephanie è di madrelingua yankee e che` **`ha dovuto`** `studiare prima la lingua italiana per potersi esprimere. [...]` NG.

Confrontando la lista degli anankastici con quella dei deontici, due considerazioni si impongono: gli esempi sono pochi, le tipologie sono tutte formali; inoltre tre testi su quattro sono normativi.

[11a], che è tratto dal Codice della strada (art. 2, comma 3) e [11c] rappresentano il tipo standard di anankastico, che si trova comunemente in testi di tipo normativo. Nel caso di [11c] le caratteristiche utili a definire "locale fumatori" un certo locale sono necessariamente possibili. Altrimenti detto: un locale potrebbe non avere quelle caratteristiche, ma se non le avesse non sarebbe un locale fumatori[24]. Il "dovere" di [11c] indica dunque una necessità anankastica, come dimostra la possibilità di sostituire il verbo con "potere", senza alcuna sostanziale variazione di significato per l'enunciato.

---

[23] Come dimostra l'uso del maiuscolo, che nelle convenzioni dei newsgroup equivale ad un tratto soprasegmentale (oralmente il locutore avrebbe alzato la voce e fatto una pausa significativa). Il fatto che il locutore abbia bisogno di sottolineare in questa maniera la predicazione di obbligo è un buon indizio sia della non universalità, sia della dipendenza di questo obbligo da una convinzione personale del locutore.

[24] Per brevità nell'esempio non sono state elencate le caratteristiche in questione.

Il caso di [11a] esemplifica un'occorrenza di anankastico comune nei testi giuridici in genere, con cui il verbo oltre a predicare una norma la pone. Le strade potrebbero, ipoteticamente, avere caratteristiche differenti da quelle elencate nel Codice, ma nella realtà chiunque voglia costruire strade in Italia deve attenersi alla norma espressa dal Codice. Questa occorrenza di "dovere" è dunque pienamente eidetico-costitutiva (Conte, A. G. 1985), perché predica un obbligo che, almeno in linea di principio, determina il modo di essere della realtà[25].

Lo stesso schema di [11c] è presente anche in [11b] ed in [11d]. [11b], secondo la definizione data in § 1.3, può essere considerato un testo normativo perché pone regole utili alla composizione di brani musicali. Il senso dell'esempio è chiaro: un brano musicale per essere buono, non può che obbedire a tre requisiti. Anche in questo caso la sostituzione di "dovere" con "potere" lascia inalterato il senso dell'enunciato.

[11d] è l'unico esempio di anankastico non proveniente da testi normativi, sempre secondo la definizione di § 1.3, di tutta la ricerca (circa 180 output analizzati), per ciò merita particolare attenzione. La sostituibilità, *salva significatione*, di "dovere" con "potere" ci assicura, ancora una volta, che a [11c], [11b] e [11d] è sottesa la stessa natura logica, tuttavia da un punto di vista pragmatico [11d] ha alcune differenze importanti rispetto a [11b] e [11c].

Mentre in [11b] e [11c] l'informazione pragmatica è completa ed indipendente dall'implicatura conversazionale[26], in [11d] l'implicatura ha un ruolo importante a causa dell'intervento di almeno due massime griceane: quantità e relazione.

Per la massima di quantità il lettore/interlocutore è in grado di capire che con "esprimersi" si intende 'esprimersi in italiano'. Stephanie è di lingua straniera ed ha studiato l'italiano per esprimersi, si suppone quindi che Stephanie sapesse già esprimersi nella sua lingua ma non nella nostra, il che ha portato all'obbligo di studiare l'italiano per esprimersi appunto in italiano. Quest'obbligo è di natura anankastica, non è infatti possibile non studiare l'italiano se ci si vuole esprimere in italiano.

Come parlanti italiano però ci sentiamo di approvare la conclusione enunciata sopra solo in base alla massima di relazione, per cui ci aspettiamo che il contributo del partner sia «appropriato alle esigenze immediate di ciascuna fase della transazione» (Grice 1989/93, p. 62). Se applichiamo la massima ad un registro formale di lingua, di cui, nonostante alcuni colloquialismi, il nostro testo è un esempio, sappiamo che "esprimersi" indica una competenza linguistica che va oltre la semplice capacità di farsi intendere, quindi una competenza per cui è necessario lo studio. La formalità del registro comunicativo, che "pesa" le parole con maggiore attenzione, ci permette dunque di classificare l'occorrenza di "dovere" in [11d] come anankastica.

Il funzionamento pragmatico di [11d] prova così due cose: la dipendenza dell'anankastico dalla completezza del contesto informativo, che deve essere "pesante"; la conseguente rarità linguistica dell'anankastico.

---

[25] Su questo tipo di anankastico, tipico in realtà non del testo giuridico *tout-court*, ma particolarmente del testo legale, o di testi aventi, anche *latu sensu*, valore di norma legale, non tornerò più in maniera diretta. Poiché però nell'analisi di [10a], che verrà proposta in § 2.3, noteremo un fenomeno simile, qualche parola val la pena spenderla. Si consideri il seguente esempio, tratto da una sentenza della Corte di Cassazione (Sezioni Penali Riunite, presidente Viola, relatore Postiglione): scrive la Corte a commento dell'ordinanaza 220/1996 emessa dalla Corte Costituzionale: «…in conclusione **deve**, dunque, affermarsi il seguente principio […]». Quest'occorrenza di "deve" potrebbe essere intesa come deontica – ci sarebbero in realtà altri modi di intendere la stessa ordinanza – ma dal punto di vista dello scrivente va intesa come anankastica, poiché con questa sentenza la Cassazione, ossia il grado supremo della magistratura ordinaria, asserisce la possibilità di intendere in uno ed in un solo modo l'ordinanza 220 della Corte Costituzionale. Si apre qui, all'interno del linguaggio giuridico, ma non solo, un'interessante punto di faglia fra interpretazione e modalità, che dovrebbe costringerci ad approfondire la rappresentazione del deontico come *subject oriented* e dell'anankastico come *object oriented.*

[26] È significativo che entrambi gli esempi siano tratti da enumerazioni, quindi da una forma comunicativa pesante e, solitamente, molto prolissa.

2.3 CONTESTI INCERTI E COOCCORRENZE. Riconsideriamo l'esempio [10a], già presentato in § 2.1:

[10a] `[...] similmente decretiamo e dichiariamo che le presenti lettere in nessun tempo` **`potranno`** `venir revocate o diminuite, ma stabili sempre e valide` **`dovranno`** `perseverare nel loro vigore [...]` NG.

Ne riprendo l'analisi, perché ciò mi permette di accennare ad alcuni problemi. Una delle differenze proposte per distinguere anankastico e deontico è la seguente: il deontico sarebbe *subject oriented*, l'anankastico invece risulterebbe *object oriented*, il primo infatti riguarderebbe un obbligo concernente un soggetto agente, il secondo definirebbe le caratteristiche costitutive di un oggetto (cfr. Conte, A. G., 1977, 1992, 1999 e Conte, M.-E., 1993, 1995).

Si può facilmente dimostrare che tutti gli esempi di deontico ed anankastico fin qui studiati rispettano questi differenti orientamenti, ma con [10a] sorge un particolare problema legato alla natura stessa della possibilità.

Si è detto che [10a] può essere assimilabile al deontico categorico perchè predica un obbligo assoluto, questa posizione sembra sostenibile se si considera l'enunciato come un divieto implicito a modificare le lettere in questione, e quindi come un obbligo imposto al soggetto grammaticale della frase: *praesentes litterae*[27].

Tuttavia se noi considerassimo l'enunciato dal punto di vista dell'estensore della bolla, allora dovremmo concludere che il dovere in questione è anankastico. Pio V infatti proclama l'impossibilità anankastica di modificare le lettere («praesentes litterae… unquam revocari aut moderari possint»), da ciò potrebbe seguire il dovere, anankastico dunque e non deontico, di preservare le lettere.

Si noti però il passaggio dalla forma passiva per esprimere l'anankasticità (*revocari aut moderari possint*), alla forma attiva per esprimere quello che non sembra più essere un obbligo imposto all'oggetto di una predicazione, ma un dovere indicato al soggetto di una possibile modificazione. In un certo senso, col passaggio dal passivo all'attivo, è come se Pio V ricorresse alla figura dell'antonomasia[28], quasi dicesse: 'bada, Messale, a rimanere sempre uguale a te: non devi modificarti'.

È dunque anankastica l'impossibilità di subire modificazioni, ma deontico il dovere di sottostare a quest'impossibilità: per ciò mi sento di confermare l'inclusione di [10a] fra i deontici.

Questa rapida analisi dell'esempio tratto dalla bolla *Quo primum tempore*[29]ci porta a fare alcune altre considerazioni sulla natura dell'intervento pragmatico nel caso dell'anankasticità.

A differenza del deontico, che sembra essere relativamente semplice da gestire nell'àmbito della comunicazione, l'anankastico risulta pragmaticamente "pesante", esso richiede, come si è già detto, contesti molto precisi, nei quali l'applicazione delle massime griceane sia poco economica, la qual cosa spiega perché l'anankastico sembra essere legato a registri formali.

Anche i registri formali però diventano problematici – è appunto il caso di [10a] – e possono talora essere, come si vedrà nell'esempio conclusivo, irrisolvibili. Questa problematica porta a chiedersi quali dunque siano i fatti pragmatici che intervengono nella determinazione dell'anankastico.

[27] C'è qui un'evidente interferenza sintattica: se ragioniamo in termini funzionalisti (cfr. Perlmuter 1983), il soggetto superficiale, "grammaticale", della frase potrebbe essere considerato l'oggetto profondo dell'enunciato. In questo senso dico che l'interpretazione deontica sembrerebbe dover essere intesa come divieto. Siamo qui in presenza di un particolare àmbito linguistico di applicazione del test "logico" proposto in § 0.1: si potrebbe dire che la forma logica del deontico vale per il soggetto superficiale perché è applicabile all'oggetto profondo, ma non è certo questa la sede adatta per una discussione sui rapporti fra sintassi e modalità.

[28] Nulla di più facile nell'elegante latino umanistico che allora anche Santa Romana Chiesa sapeva parlare.

[29] Si sarebbero dovute spendere più parole sui fatti di traduzione.

A questo riguardo qui posso solo avanzare alcune idee, che faranno da bussola a successive ricerche. Un primo fatto mi sembra inoppugnabile: l'anankastico neutralizza la massima di quantità, poiché richiede che tutte le informazioni siano espresse, limitando così l'azione della massima di quantità a sottintendere solo le informazioni banali, ossia quelle che possono certamente ed univocamente essere ricavate dal contesto[30]. A questa prolissità dell'anankastico mi sembra facciano da *pendant* altri fatti: accennerò qui solo a due.

Per la massima di relazione ci aspettiamo che: «il contributo del partner sia appropriato alle esigenze immediate di ciascuna fase della transazione» (Grice 1989, tr. it. p. 62); quest'aspettativa, soprattutto nel parlato e nei registri informali, va contro l'anankastico. Si consideri il seguente esempio, normativo ma informale:

[12] `[...] da Pinchirri devi avere la cravatta [...] e da me dovresti avere la cultura e la sensibilità! [...]` NC.

In questo caso è adeguato alle "esigenze immediate della transazione" solo indicare l'esistenza di requisiti necessari per la frequentazione di Pinchi[o]rri[31] e del locale dello scrivente, se poi tali requisiti siano predicabili come anankastici o come deontici non lo sapremo mai, non è infatti adeguato a quest'atto linguistico darci informazioni a proposito.

Si noti però che applicando il test logico di § 0.1 è più facile intendere le due occorrenze di "dovere" come deontici, nulla infatti sembra impedire la possibilità di andare da Pinchi[o]rri senza cravatta o di frequentare il locale dello scrivente non avendo la necessaria cultura. "Dovresti avere la cultura e la sensibilità" ha però buone carte per essere considerato un anankastico, potrebbe infatti essere accettabilmente parafrasato 'per essere veramente considerato mio cliente devi avere cultura e sensibilità, altrimenti puoi anche venire a mangiare da me ma non avrai i requisiti per essermi cliente'. Non è esente da sfumature anankastiche nemmeno "da Pinchi[o]rri devi avere la cravatta": si potrebbe infatti intendere la cosa come un requisito necessario per essere veramente considerati clienti di Pinchi[o]rri, e non come un obbligo.

Una considerazione va qui fatta: è concezione comune, anche nella cultura giuridica[32], che la nozione di "dovere" comporti una conseguenza negativa per il soggetto, del tipo: o fai così o sei punito. Il fatto che a [12] possa applicarsi la formula logica del deontico smentisce questa idea del "dovere": certamente ci sono deontici meno forti e deontici più forti, ci sono addirittura deontici fortissimi, che vincolano il soggetto ad una scelta morale alta (tipicamente tali sono i deontici categorici), ma linguisticamente non pare né utile né verosimile determinare la differenza deontico / anankastico partendo dall'idea che il primo, a differenza del secondo, abbia un supposto valore punitivo.

È comunque innegabile che in assenza di informazione sufficiente per determinare se le occorrenze di "dovere" in [12] siano anankastiche o deontiche, la lettura deontica sembra più immediata e, generalmente, meglio adeguata alla situazione comunicativa.

Il primo fatto che dunque denuncia la maggior difficoltà linguistica dell'anankastico sembra così riguardare la massima di relazione; il secondo ha invece a che fare non direttamente con le massime, ma con una caratteristica propria della pragmatica dei registri formali.

Nei registri informali, e naturalmente nel parlato, il livello pragmatico ha un ruolo comunicativo chiaro, vuoi perché i partecipanti alla comunicazione non hanno generalmente interesse a giocare sulle anfibologie, sui sottintesi, sui "trucchi" della comunicazione[33], vuoi perché l'estensione e l'organizzazione del testo è più semplice e limitata, avendo così minori possibilità di interferire col livello pragmatico.

---

[30] Mi pare che l'esempio [11d] sia una buona riprova di ciò: in quel caso sarebbe stato inutile dire "esprimersi in italiano" perché l'informazione era ricavabile, senza alcun problema, dal contesto.

[31] Che in realtà si tratti della celebre Enoteca Pinchiorri di Firenze non pare da dubitarsi [N. di M.B.].

[32] Devo l'informazione a Paolo Di Lucia.

[33] Non a caso l'umorismo appartiene sempre ad un livello formale.

I registri formali contravvengono sempre alla seconda delle due caratteristiche appena elencate e spesso anche alla prima, questo fatto è esiziale sia per l'anankastico, che certamente, data la sua "pesantezza" comunicativa, non può permettersi ambiguità, sia per la possibilità stessa di determinare se un'occorrenza sia deontica od anankastica. Siamo così in presenza di quello che potrei definire il paradosso dell'anankastico: l'anankastico può vivere quasi esclusivamente nei registri formali, ma se tali registri non disambiguano, l'anankastico muore perché viene meno anche la possibilità di determinare la deonticità di un'occorrenza di "dovere". Altrimenti detto, se di un'occorrenza di "dovere" in un registro formale non siamo in grado di dire se sia deontica od anankastica, allora quell'occorrenza è indecidibile[34].

Porto a riprova il seguente esempio:

[13] `[...] Fra parentesi indicheremo i testi biblici ai quali si fa riferimento, che dovrebbero essere pazientemente cercati e letti per una piena intelligenza delle cose dette [...]` NG.

In questo caso la ricerca e la lettura dei testi biblici è un dovere deontico, che può anche essere eluso, come parrebbe suggerire l'uso del condizionale, od è un obbligo anankastico non eludibile, come invece indurrebbe a pensare la finale implicita?

Ritengo che si potrebbero portare argomenti validi a sostegno dell'una e dell'altra ipotesi, ma penso anche che proprio la validità degli argomenti pro deontico e di quelli pro anankastico, e la conseguente validità della reciproca confutazione, dimostri come il registro formale comprometta, in questo ed in altri simili casi, la possibilità di scelta fra anankastico e deontico proprio per le ragioni accennate sopra[35].

3. CONCLUSIONI. Alla fine del mio intervento sulla modalità anankastica e deontica di "dovere", posso dire che queste pagine più che un articolo sono un *memorandum* delle cose da fare, delle direzioni che credo la ricerca debba prendere.

Per ciò mi piace finire non con un riassunto degli argomenti, né, tanto meno, con una proposta, ma col doveroso riconoscimento di un debito. Le mie proposte camminano nel solco tracciato da Amedeo G. Conte in filosofia del diritto e da Maria-Elisabeth Conte in linguistica testuale; da linguista però non posso non deprecare, soprattutto a confronto con la vivacità del dibattito in àmbito di filosofia del diritto, che il sentiero additato da Maria-Elisabeth Conte sia stato ancora così poco battuto, spero dunque, *si parva licet*, di essermi addentrato un poco in quella direzione.

---

[34] Ancora una volta sono costretto a chiedere venia per l'apoditticità delle mie affermazioni, che spero servano almeno da guida al lettore.

[35] Non a caso nella prosa formale si tende a fare un certo uso di avverbi modali (cfr. Venier 1991), come "necessariamente". Si consideri il seguente passo: «al riconoscimento giuridico deve necessariamente seguire la volontà di dare attuazione ai diritti culturali e devono concorrere le condizioni storiche ed economiche affinché la buona volontà dei governanti possa dare spessore e corpo al diritto formale» (A). L'avverbio "necessariamente" serve all'autore per indicare che, per lui, l'occorrenza di dovere è anankastica; se ora noi riscrivessimo il passo eliminando l'avverbio, ci troveremmo nelle condizioni di 26, non avremmo cioè elementi per decidere se "dovere" sia anankastico o deontico.

BIBLIOGRAFIA.

AA. VV.
1904 *Missale Romanum ex decreto Sacrosanti Concilii Tridentini restitutum, S. Pii V Pontificis Maximi jussu editum, Clementis VIII, Urbani VIII et Leonis XIII auctoritate recognitum,* editio secundam post alteram uti typicam a S.R.C. declaratam, Ratisbonae – Romae - Neo Eboraci - Cincinnati, Sumptibus chartis et typis Friederici Pustet, 1904.

BAZZANELLA
1994 Carla Bazzanella, *Le facce del parlare: un approccio pragmatico all'italiano parlato*, Scandicci, La nuova Italia, 1994.

BIERKELUND - BOYSENE - KJÆRSGAARD
2003 *Aspects de la modalité*, edité par Marete Bierkelund, Gerhard Boysene et Poul Søren Kjærsgaard, Tübingen, Max Niemeyer Verlag, 2003 "Linguistische Arbeiten".

CONTE, A. G.
1977 Amedeo Giovanni Conte, *Aspetti della semantica del linguaggio deontico*, in DI BERNARDO 1977, pp. 147-165.
1985 Amedeo Giovanni Conte, *Regole eidetico-costitutive*, in "Nuova civiltà delle macchine" III-IV (1985) 26-33.
1986 Amedeo Giovanni Conte, *Fenomeni di fenomeni*, in GALLI 1986, pp. 167-198.
1992 Amedeo Giovanni Conte, *Deontica aristotelica*, in "Rivista internazionale di filosofia del diritto" LIX (1992) 178-252.
1999 Amedeo Giovanni Conte, *Three Levels of Deontics*, in EGIDI 1999, pp. 205-214.

CONTE, M.-E.
1993 Maria-Elisabeth Conte, *Modalità fra semantica e pragmatica*, in NEGRI - POLI 1993, pp. 139-151
1995 Maria-Elisabeth Conte, *Epistemico, deontico, anankastico*, in GIACALONE RAMAT - CROCCO GALÈAS 1994, pp. 3-9.
1999/88 Maria-Elisabeth Conte, *Condizioni di coerenza*, Alessandria, Edizioni dell'Orso, 1999. Nuova edizione, con l'aggiunta di due saggi a cura di Bice Mortara Garavelli, di Maria-Elisabeth Conte, *Condizioni di coerenza. Ricerche di linguistica testuale*, Firenze, La Nuova Italia Editrice, 1988 "Pubblicazioni della Facoltà di Lettere e filosofia dell'Università di Pavia" 46.

DENDALE - TASMOWSKI
2001 *Le conditionnel en Français*, édité par Patrik Dendale et Liliane Tasmowski, Metz, Université de Metz, 2001 "Recherches linguistiques" 25.

DI BERNARDO
1977 *Logica deontica e semantica*, a cura di Giuseppe Di Bernardo, Bologna, Il Mulino, 1977.

DI LUCIA
2000 Paolo Di Lucia, *"Sollen" in Herbert Spiegelberg*, in VERONESI 2000, pp. 69-84.

EGIDI
1999 *In Search of a New Humanism: the Philosophy of Georg Henrik von Wright*, edited by Rosaria Egidi, Boston - Dordrecht - London, Kluwer Academic Press, 1999 "Synthèse Library".

ERNOUT - THOMAS
1953 Alfred Ernout - François Thomas, *Syntaxe latine*, Paris, Librairie Klincksieck, 1953².

GALLI
1986 *Interpretazione ed epistemologia. Atti del VII colloquio internazionale sulla interpretazione. Macerata 25-27 Marzo 1985*, a cura di Giuseppe Galli, Torino, Marietti, 1986.

GHISLIERI → PIO V.

GIACALONE RAMAT - CROCCO GALÈAS
1994 *From Pragmatics to Syntax. Modality in Second Language Aquisition*, edited by Anna Giacalone Ramat and Grazia Crocco Galèas, Tübingen, Gunter Narr Verlag, 1994.

GRICE
1989/93 Paul Grice, *Logica e conversazione. Saggi su intenzione, significato e comunicazione*, traduzione italiana di Giorgio Moro, Bologna, il Mulino, 1993 [edizione originale: Paul Grice, *Studies in the way of words*, Cambridge (Mass.) - London, Harvard University Press, 1989].

HEINE
1992 Bernard Heine, *On the Nature of Semantic Change in Grammaticalization*, in NEGRI - POLI 1993, pp. 11-28.

KIEFER
1987 Kiefer Ferenc, *On Defining Modality*, in "Folia linguistica" XXI (1992) 67-94.

KRONNING
2001 Hans Kronning, *Au-delà du déontique et de l'épistémique*, in PRANDI - RAMAT 2001, pp. 97-117.
2001 Hans Kronning, *Nécessité et hypothèse: "devoir" non déontique au conditionnel*, in DENDALE - TASMOWSKI 2001, pp. 251-76.
2003 Hans Kronning, *Modalité et évidentialité*, in BIERKELUND - BOYSENE - KJÆRSGAARD 2003, pp. 131-51.

LEVINSON
1983 Stephen C. Levinson, *Pragmatics*, Cambridge, Cambridge University Press, 1983 "Cambridge textbooks in linguistics".

MORTARA GARAVELLI
2001 Bice Mortara Garavelli, *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, Torino, Einaudi, 2001, "Piccola biblioteca Einaudi".

NEGRI - POLI
1993 *La semantica in prospettiva diacronica e sincronica. Atti del convegno della Società italiana di Glottologia (Macerata e Recanati, 22-24 Ottobre 1992)*, a cura di Mario Negri e Diego Poli, Pisa, Giardini editori e stampatori, 1993.

PERLMUTTER
1983 David M. Perlmutter, *Personal vs. Impersonal Constructions*, in "Natural Language and Linguistic Theory" I (1983) 140-200.

PIO V
1570/1904 [Antonio Michele Ghislieri, papa Pio V (santo)], *Pius Episcopus Servus Servorum Dei, ad perpetuam rei memoriam. Quo primum tempore ad Apostolatus apicem [...]*, in AA. VV.1904, p. 1.

PRANDI
2001 Michele Prandi, *Maria-Elisabeth Conte: tra semiotica e linguistica*, in PRANDI - RAMAT 2001, pp. 13-26.

PRANDI - RAMAT
2001 *Semiotica e linguistica. Per ricordare Maria Elisabeth Conte*, a cura di Michele Prandi e Paolo Ramat, Milano, Franco Angeli Editore, 2001, “Materiali Linguistici dell’Università di Pavia” 32.

ROVERE
2005 Giovanni Rovere, *Capitoli di linguistica giuridica. Ricerche su* corpora *elettronici*, Alessandria, Edizioni dell’Orso, 2005 “Gli argomenti umani” 9.

STALNAKER
1970 Robert C. Stalnaker, *Pragmatics*, in “Synthèse” XXII (1970) 512-530.

TOGNINI-BONELLI
2001 Elena Tognini-Bonelli, *Corpus Linguistics at Work*, Amsterdam - Philadelphia, John Benjamins Publishing Company, 2001 “Studies in Corpus Linguistics” 6.

VENIER
1991 Federica Venier, *La modalizzazione assertiva. Avverbi modali e verbi parentetici*, Milano, Franco Angeli Editore, 1991 “Materiali Linguistici dell’Università di Pavia” 5.

VERONESI
2000 *Linguistica giuridica italiana e tedesca – Rechtslinguistik des Deutschen und Italienischen*, a cura di Paola Veronesi, Padova, Unipress, 2000.

WRIGHT
1963 Georg Henrik von Wright, *The Logic of Preference*, Edinbourgh, Edinbourgh University Press, 1963.
1963b Georg Henrik von Wright, *Norm and Action, a Logical Enquiry*, London, Routledge & Kegan Paul, 1963 “International Library of Philosophy and Scientific Method”.

CORPORA DI RIFERIMENTO.

Athenaeum Corpus	http://www.bmanuel.org/projects/at-HOME.html
bmanuel.org	http://www.bmanuel.org
corpora.unito.it	http://www.corpora.unito.it/
Jus Jurium	http://www.bmanuel.org/projects/ju-HOME.html
NUNC-IT Generic	http://www.bmanuel.org/projects/ng-HOME.html
NUNC-IT Cooking	http://www.bmanuel.org/projects/ng-HOME.html
NUNC-IT Photo	http://www.bmanuel.org/projects/ng-HOME.html
NUNC-IT Motor	http://www.bmanuel.org/projects/ng-HOME.html

AMEDEO GIOVANNI CONTE
(Accademia Nazionale dei Lincei)

22. Valori non-normativi di verbi deontici in testi normativi.

Τὸ δ'ἀτρεκὲς ἐν βαθεῖ ἐστι.
Oracolo caldaico, fr. 182[1].

Die Tiefe muß man verstecken.
Wo?
An der Oberfläche.
Hugo von Hofmannsthal, *Buch der Freunde*[2].

SOMMARIO. **0.** Introduzione. **1.** Valore non-normativo di verbi deontici in testi *non-normativi*. **2.** Valore non-normativo di verbi deontici in testi *normativi*.

0. INTRODUZIONE. Il saggio *Valori non-normativi di verbi deontici in testi normativi* indaga il valore dei verbi *deontici* nei testi *normativi*.

0.1 DEFINIZIONE. Chiamo verbi deontici i verbi modali tra i cui valori vi sia un valore deontico (ossia i verbi *prima facie* deontici): ad esempio, *dovere* e *potere* in italiano; *pouvoir* e *devoir* in francese; *sollen*, *müssen*, *dürfen*, *können* in tedesco; *ought to*, *must*, *can*, *may* in inglese.

0.1.1 *'DOVERE' > 'POTERE'. Per la sua singolare vicenda semantica, spicca, tra questi verbi, il tedesco *dürfen*.

Il verbo deontico tedesco *dürfen* (il cui senso primo e primario, nel tedesco odierno, il *Neuhochdeutsch* [nuovo alto tedesco], è 'potere', 'avere il permesso di'[3]) originariamente significava (non: 'potere', ma) 'dovere'.

0.1.2 *'DOVERE' > 'DOVERE'. Il senso originario di 'dovere' (oggi scomparso nel tedesco *dürfen*) permane, invece, in numerosi verbi, di altre lingue indoeuropee, che a *dürfen* sono etimologicamente affini.

Ecco sei esempi: due esempi sono desunti da due lingue (estinte) *germaniche* (gotico, antico nordico [norréno]); tre esempi sono desunti da tre lingue *slave* (russo, polacco, ceko); un esempio è desunto da una lingua *romanza* (il romeno).

(j) Lingue *germaniche*:
[1a] gotico: *thaurban* ('avere bisogno di', '*bedürfen*');
[1b] antico nordico (norréno): *Þarfa* ('essere necessario', '*nötig sein*'),
Þarfna ('avere bisogno di', '*bedürfen*') de Vries 1962.

(ij) Lingue *slave*:
[2a] russo: *требоваться* "*trébovat'sja*" ('occorrere', 'essere necessario', '*erforderlich sein*', '*nötig sein*').

1 'L'evidente è nel profondo'; ed.: *Oracoli caldaici*, a cura di Angelo Tonelli, Milano, Rizzoli, 1995, p. 208.
2 'La profondità va nascosta. Dove? Alla superficie'; ed.: *Buch der Freunde*, herausgegeben von Ernst Zinn, Frankfurt am Main, Insel-Verlag, 1965, p. 51.
3 Esempio: *Darf ich?* significa 'Posso?', '*May I?*'.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 363-370.

[2b] polacco: *trzeba* (‘bisogna’; ‘*es ist nötig*’; ‘*il faut*’);
[2c] ceko: *třeba* (‘è necessario’, ‘*es ist nötig*’);

(iij) Lingue *romanze*:
[3] romeno: *a trebui* (‘dovere’)[4].

0.2. LIMITI. Annuncio sùbito due limiti della mia indagine.

0.2.1 PRIMO LIMITE. *In primo luogo*, la mia indagine si limita a *verbi* deontici (ad esempio: *dovere*, *potere*): essa non tratta *altri* termini deontici: in particolare, *aggettivi* (ad esempio: *obbligatorio*, *permesso*) e *sostantivi* (ad esempio: *obbligo*, *permesso*).

Recentissimamente ho scoperto[5] che, quasi sei secoli prima che Georg Henrik von Wright [Helsinki (in svedese: Helsingfors), 14 giugno 1916 - Helsinki, 16 giugno 2003] fondasse la logica deontica[6], era apparsa una enumerazione degli *aggettivi* deontici arabi. Di questa enumerazione è autore al-Malik al-Afḍal in un libro, in arabo, del 1370: *Nuzhat aẓ-ẓurafā' wa tuḥfat al-Hulafā'* [*Svago per gli uomini raffinati e dono per i califfi*][7].

Gli aggettivi deontici arabi sono, secondo al-Malik al-Afḍal, cinque:

[4a] *wāǧib* ‘obbligatorio’;
[4b] *mandūb* ‘raccomandabile’;
[4c] *muḥarram* ‘vietato’[8];
[4d] *makrūh* ‘riprovevole’;
[4e] *mubaḥ* ‘permesso’.

0.2.2 SECONDO LIMITE. *In secondo luogo*, la mia indagine si limita a 12 verbi deontici di quattro lingue: in particolare, a 12 verbi deontici delle quattro lingue (tedesco, italiano, francese, retoromanico[9]) dei testi legislativi svizzeri: *dastgar*, *devoir*, *dovere*, *duair*, *dürfen*, *können*, *müssen*, *potere*, *pouvoir*, *pudair*, *sollen*, *stuair*[10].

---

[4] Cfr. Conte 2007a.
[5] Mia fonte: Renato Tràini [*1923].
[6] Wright 1951.
[7] Edizione critica con versione italiana annotata: Tràini 2005; cfr. anche Tràini 2006. Sui modi deontici in al-Malik al-Afḍal, cfr. Conte 2006a.
[8] Presumo che *muḥarram* ‘vietato’ sia etimologicamente affine al termine arabo *harām*, *harīm* ‘vietato’, ‘inviolabile’; termine che (con la mediazione del turco *harem*) è entrato in italiano (nella forma *harem*) come designazione del ginecéo, ossia della parte della casa musulmana riservata alle donne, parte alla quale era vietato (proibito, interdetto) l’accesso agli estranei.

Il caso di *harem* è un fenomeno filosoficamente provocante: un termine deontico (un deontónimo), e precisamente *harem* ‘vietato’, funge non da termine *qualificativo*, ma da termine *designativo*. Mi riferisco al paradigma diadico: termine *qualificativo* vs. termine *designativo*, paradigma concepito da Uberto Scarpelli [1924-1993] e fecondamente ripreso da Giuseppe Lorini [*1969]. Analogo, in latino, il rapporto intercorrente tra l’aggettivo (un axiónimo) *incestus* ‘impuro’, ‘*unkeusch*’, e due nomi dell’incesto (dello *Inzest*, della *Unzucht*): il sostantivo neutro *incestum, incesti* (II. declinazione) ed il sostantivo maschile *incestūs, incestūs* (IV. declinazione).

(Un *curiosum* filosoficamente irrilevante: in Sicilia, ho trovato un toponimo omonimo di un deontónimo: *Divieto*. Divieto è un paese in provincia di Messina.)

[9] Il retoromanico (*vel* romancio) è una delle tre lingue romanze (neolatine) della Svizzera. (Le altre due sono l’italiano ed il francese.) È parlato nel Cantone dei Grigioni [*Grischun* in retomanico; *Graubünden* in tedesco; *Grisons* in francese]. Il capoluogo dei Grigioni è Coira [*Cuira* in retomanico; *Chur* in tedesco; *Coire* in francese]; ma il luogo più noto al filosofo è Sils Maria (il villaggio ove Friedrich Wilhelm Nietzsche [1844-1900] trascorse l’estate tra il 1881 ed il 1889).

Alcuni glottónimi che designano il retoromanico (romancio) sono: in retoromanico: *retorumantsch*, *rumantsch*; in tedesco: *Rätoromanisch*, *Romaunsch*, *Romauntsch*, *Rumantsch*, *Rumauntsch*; in francese: *rhéto-roman*, *romanche*, *roumanche*.

[10] I verbi deontici, se sono *in suppositione materiali*, sono posti in corsivo *senza* virgolette. Il senso dei verbi deontici è inscritto tra virgolette semplici: ‘ ’.

(j) ***Tedesco***:

[5a] *dürfen* [italiano: 'potere'; francese: '*pouvoir*'; retoromanico: '*dastgar*', '*pudair*'],
[5b] *können* [italiano: 'potere'; francese: '*pouvoir*'; retoromanico: '*dastgar*', '*pudair*'],
[5c] *müssen* [italiano: 'dovere'; francese: '*devoir*'; retoromanico: '*duair*', '*stuair*'],
[5d] *sollen* [italiano: 'dovere'; francese: '*devoir*'; retoromanico: '*duair*', '*stuair*'].

(ij) ***Italiano***:

[6a] *dovere* [tedesco: '*sollen*', '*müssen*'; francese: '*devoir*'; retoromanico: '*duair*', '*stuair*'],
[6b] *potere* [tedesco: '*dürfen*', '*können*'; francese: '*pouvoir*'; retoromanico: '*dastgar*', '*pudair*'].

(iij) ***Francese***:

[7a] *devoir* [tedesco: '*sollen*', '*müssen*'; italiano: 'dovere'; retoromanico: '*duair*', '*stuair*'],
[7b] *pouvoir* [tedesco: '*dürfen*', '*können*'; italiano: 'potere'; retoromanico: '*dastgar*', '*pudair*'].

(iiij) ***Retoromanico***:

[8a] *dastgar* [tedesco: '*dürfen*', '*können*'; italiano: 'potere'; francese: '*pouvoir*'],
[8b] *duair* [tedesco: '*sollen*', '*müssen*'; italiano: 'dovere'; francese: '*devoir*'],
[8c] *pudair* [tedesco: '*dürfen*', '*können*'; italiano: 'potere'; francese: '*pouvoir*'],
[8d] *stuair* [tedesco: '*sollen*', '*müssen*'; italiano: 'dovere'; francese: '*devoir*'].

0.3 I MATERIALI DEL PRESENTE SAGGIO. Di questi 12 verbi deontici (*dastgar*, *devoir*, *dovere*, *duair*, *dürfen*, *können*, *müssen*, *potere*, *pouvoir*, *pudair*, *sollen*, *stuair*) ho indagato il valore in 46 documenti tratti

(j) dal Codice civile svizzero (nelle sue tre lingue: tedesco, italiano, francese; i tre testi sinottici si intitolano: *Schweizerisches Zivilgesetzbuch*, *Codice civile svizzero*, *Code civil suisse*);

(ij) dalla Costituzione federale della Confederazione Svizzera (nelle sue quattro lingue: tedesco, italiano, francese, retoromanico; i quattro testi sinottici si intitolano: *Bundesverfassung der Schweizerischen Eidgenossenschaft*, *Costituzione federale della Confederazione Svizzera*, *Constitution fédérale de la Confédération Suisse*, *La Nova Constituziun federala*)[11].

In particolare, ho esaminato:

(j) 31 ricorrenze di otto verbi deontici, appartenenti a tre lingue (tedesco, italiano, francese): *devoir*, *dovere*, *dürfen*, *können*, *müssen*, *potere*, *pouvoir*, *sollen*, in 30 documenti tratti dal Codice civile svizzero nelle sue tre lingue (*Schweizerisches Zivilgesetzbuch*, *Codice civile svizzero*, *Code civil suisse*);

(ij) 19 ricorrenze di 11 verbi deontici, appartenenti a quattro lingue (tedesco, italiano, francese, retoromanico): *dastgar*, *devoir*, *dovere*, *duair*, *dürfen*, *können*, *potere*, *pouvoir*, *pudair*, *sollen*, *stuair*, in 16 documenti tratti dalla Costituzione federale della Confederazione Svizzera nelle sue quattro lingue (*Bundesverfassung der Schweizerischen Eidgenossenschaft*, *Costituzione federale della Confederazione Svizzera*, *Constitution fédérale de la Confédération Suisse*, *La nova Constituziun federala*).

---

[11] Enumero (in ordine alfabetico) i sette testi normativi della mia *base testuale*:

(j) *Bundesverfassung der Schweizerischen Eidgenossenschaft*;
(ij) *Code civil suisse*;
(iij) *Codice civile svizzero*;
(iiij) *Constitution fédérale de la Confédération Suisse*;
(v) *Costituzione federale della Confederazione Svizzera*;
(vj) *La Nova Constituziun federala*;
(vij) *Schweizerisches Zivilgesetzbuch*.

I 46 documenti da me raccolti *non* sono riprodotti nel presente saggio *Valori non-normativi di verbi deontici in testi normativi*. Essi sono éditi nel saggio: Amedeo G. Conte, *Fenomeni normativi. Un'indagine non-filosofica*[12].

0.4. LA DOMANDA. Vengo alla domanda. Può un verbo *deontico* avere valore *non-normativo* entro un testo *normativo*? Alla risposta sono dedicati il § 1 ed il § 2.

1. VALORE NON-NORMATIVO DI VERBI DEONTICI IN TESTI *NON-NORMATIVI*. *Sembra* ovvio che, in un testo *non-normativo*, un verbo *deontico* possa avere valore *non-normativo.*

1.1 IL CASO DI *TRACTATUS* 7. Consideriamo, ad esempio, un celebre testo *non-normativo*: il *Tractatus logico-philosophicus*, 1921, di Ludwig Wittgenstein [Wien 1889-Cambridge 1951]. In particolare, consideriamo *Tractatus* 7:

[9a] *Wovon man nicht sprechen kann, darüber muß man schweigen*.
[9b] 'Su ciò, di cui non si può parlare, si deve tacere'.

La settima parola di *Tractatus* 7 è la forma verbale *muß* (terza persona singolare dell'indicativo presente del verbo *müssen*)[13]. Ora, questa settima parola, *muß*, di *Tractatus* 7, è alternamente interpretata

(j) ora in senso *normativo* ('è doveroso'),
(ij) ora in senso *non-normativo* ('non si può non')[14].

1.2 LE TRADUZIONI DI *TRACTATUS* 7. L'alternanza delle due opposte interpretazioni (interpretazione in senso *normativo*, interpretazione in senso *non-normativo*) è mostrata dalle seguenti quindici traduzioni (che enumero in ordine cronologico) di *Tractatus* 7.

[10a] *Whereof one cannot speak, thereof one must be silent*.
1922: trad. inglese ascritta a Frank Plumpton Ramsey e Charles Kay Ogden.
[10b] *De lo que no se puede hablar, mejor es callarse*.
1957: trad. castigliana di Enrique Tierno Galván.
[10c] *O čemu se ne može govoriti, o tome se mora šutjeti*.
1960: trad. croata di Gajo Petrović.
[10d] *Ce dont on ne peut parler, il faut le taire*. 1961: trad. francese di Pierre Klossowski.
[10e] *What we cannot speak about we must consign to silence*.
1961: trad. inglese di David F. Pears e Brian [B. F.] McGuinness.
[10f] *Vad man icke kan tala om, därom måste man tiga*.
1962: trad. svedese di Anders Wedberg.
[10g] *O czym nie można mówić, o tym trzeba milczeć*.
1970: trad. polacca di Bogusław Wolniewicz.
[10h] *Mistä ei voi puhua, siitä on vaiettava*. 1971: trad. finnica di Heikki Nyman.
[10i] *What we cannot speak about we must pass over in silence*.
1971: trad. inglese di David F. Pears e Brian [B. F.] McGuinness.
[10j] *O čemer ne moremo govoriti, o tem moramo molčati*.
1976: trad. slovena di Frane Jerman.

---

[12] Conte 2007b.
[13] Al tedesco *müssen* sono etimologicamente affini l'inglese *must*, il nederlandese *moeten*, lo svedese *måste*.
[14] In molte traduzioni, il dilemma ermeneutico (senso *normativo*, o senso *non-normativo*?) è eluso con la scelta d'un lessema ancipite: ancìpite, come è ancìpite il tedesco *muß* (ad esempio: italiano *deve*, inglese *must*, svedese *måste*).

[10k] *Γιά ὅσα δέν μπορεῖ νά μιλάει κανείς, γιά αὐτά πρέπει νά σωπαίνει.*
1978: trad. neogreca di Θανάσης Κιτσόπουλος.

[10l] *De lo que no se puede hablar hay que callar.*
1987: trad. castigliana di Jacobo Muñoz e Isidoro Reguera.

[10m] *Acerca daquilo de que se não pode falar, tem que se ficar em silêncio.*
1987: trad. portoghese di Manuel Santos Lourenço.

[10n] *Do que se não pode falar, é melhor calar-se.*
1987: trad. portoghese di José Tiago Fonseca de Oliveira.

[10o] *Mintza ezin daitekeenari buruz isildu egin behar da*[15].
1990: trad. basca (euskara) di José Luis Álvarez Santa Cristina.

2. VALORE NON-NORMATIVO DI VERBI DEONTICI IN TESTI *NORMATIVI*. *Sembra* ovvio che, entro un testo *non-normativo* (come il *Tractatus*) un verbo *deontico* possa alternamente avere sia valore *normativo*, sia valore *non-normativo*.

Ma *non* è ovvio che questa possibilità sussista anche nell'ipotesi che un verbo *deontico* ricorra in un testo *normativo*. Tuttavia, questa paradossale possibilità sussiste: in un testo *normativo*, un verbo *deontico* può avere

(j) *non solo* valore *normativo*[16],
(ij) *ma anche* valore *non-normativo* (§§ 2.1-5).

La possibilità che, in un testo *normativo*, un verbo *deontico* abbia valore *non-normativo* è mostrata da cinque esempi:

(j) un esempio di *sollen* con valore *non-normativo* (§ 2.1);
(ij) un esempio di *müssen* con valore *non-normativo* (§ 2.2);
(iij) un esempio di *dürfen* con valore *non-normativo* (§ 2.3);
(iiij) un (primo) esempio di *können* con valore *non-normativo* (§ 2.4);
(v) un (secondo) esempio di *können* con valore *non-normativo* (§ 2.5)[17].

Ognuno di questi cinque esempi è un *exemplum contrarium* [controesempio, *Gegenbeispiel, counterexample*] il quale falsifica la tesi (apparentemente intuitiva) secondo la quale un verbo *deontico*, se appare nel contesto d'un testo *normativo*, partécipa (è partécipe) della *normatività* del testo *normativo* nel quale esso appare, e perciò stesso ha (entro quel testo) valore *normativo* (tesi della μέθεξις).

---

[15] In *Tractatus* 7, appare anche un altro verbo deontico: *können*, per il quale si ripropone il dilemma ermeneutico (senso *normativo*, o senso *non-normativo*?) che sussiste per *müssen*.

[16] Un esempio di valore *normativo* d'un verbo *deontico* in un testo *normativo* è il valore *normativo* del verbo *deontico* tedesco *sollen* nel testo tedesco del *Codice civile svizzero* (*Schweizerisches Zivilgesetzbuch*):

[11a] *Kann dem Gesetz keine Vorschrift entnommen werden, so* soll *das Gericht nach Gewohnheitsrecht und, wo auch ein solches fehlt, nach der Regel* entscheiden*, die es als Gesetzgeber aufstellen würde.*
Schweizerisches Zivilgesetzbuch, art. 1, comma 2.

Sia nel corrispondente testo italiano, sia nel corrispondente testo francese, *non* ricorrono verbi deontici:

[11b] Nei casi non previsti dalla legge il giudice decide secondo la consuetudine e, in difetto di questa, secondo la regola che egli adotterebbe come legislatore. *Codice civile svizzero*, art. 1, comma 2.

[11c] *À défaut d'une disposition légale applicable, le juge prononce selon le droit coutumier et, à défaut d'une coutume, selon les règles qu'il établirait s'il avait à faire acte de législateur.*
Code civil suisse, art. 1, comma 2.

[17] Per semplicità e brevità, limito la mia ricerca di *exempla contraria* ai verbi deontici tedeschi.

2.1 PRIMO *EXEMPLUM CONTRARIUM*. Valore *non-normativo* del verbo *deontico* tedesco *sollen* in un testo *normativo*.

[12a] Beschließt die Bundesversammlung einen Gegenentwurf, so werden den Stimmberechtigten auf dem gleichen Stimmzettel drei Fragen vorgelegt.
Jeder Stimmberechtigte kann erklären, [...] welche der beiden Vorlagen in Kraft treten soll, falls Volk und Stände beide Vorlagen dem geltenden Recht vorziehen **SOLLTEN**[18] [italiano: ∅; francese: ∅; retoromanico: ***DUESSAN***].
Bundesverfassung der Schweizerischen Eidgenossenschaft, *Schlußbestimmungen des Bundesbeschlusses vom 18. Dezember 1998*;

[12b] Italiano:
Se l'Assemblea federale adotta un controprogetto, ai votanti sono poste sulla stessa scheda tre domande.
Ogni votante può dichiarare [...] quale dei due testi dovrà entrare in vigore nel caso in cui Popolo e Cantoni li abbiano preferiti entrambi al diritto vigente.
Costituzione federale della Confederazione Svizzera, *ibidem*,

[12c] Francese:
Lorsque l'Assemblée fédérale élabore un contre-projet, trois questions seront soumises aux électeurs sur le même bulletin de vote.
Chaque électeur peut déclarer [...] lequel des deux textes devrait entrer en vigueur au cas où le peuple et les cantons préféreraient les deux textes au régime en vigueur.
Constitution fédérale de la Confédération Suisse, *ibidem*,

[12d] Retoromanico:
Decida l'assamblea federala in cuntraproject, vegnan preschentads als votants sin il medem cedel da vuschar trais dumondas.
Mintga votant po declerar [...] tgenin dals dus projects che duai entrar en vigur, sche pievel e chantuns **DUESSAN**[19] [tedesco: ***SOLLTEN***; italiano: ∅; francese: ∅**]** dar la preferenza a domadus projects avant il dretg en vigur.
Nova Constituziun federala, *ibidem*.

2.2 SECONDO *EXEMPLUM CONTRARIUM*. Valore *non-normativo* del verbo *deontico* tedesco *müssen* in un testo *normativo*.

[13a] Wer glaubhaft macht, daß er in seiner Persönlichkeit widerrechtlich verletzt ist oder eine solche Verletzung befürchten **MUSS**[20] [italiano: ∅; francese: ∅] und daß ihm aus der Verletzung ein nicht leicht wiedergutzumachender Nachteil droht, kann die Anordnung vorsorglicher Maßnahmen verlangen.
Schweizerisches Zivilgesetzbuch, art. 28c, comma 1;

[13b] Italiano:
Chi rende verosimile una lesione illecita alla sua personalità, imminente o attuale e tale da potergli causare un pregiudizio difficilmente riparabile, può chiedere al giudice di ordinare provvedimenti cautelari. *Codice civile svizzero*, *ibidem*,

[13c] Francese:
Celui qui rend vraisemblable qu'il est objet d'une atteinte illicite, imminente ou actuelle, et que cette atteinte risque de lui causer un préjudice difficilement réparable, peut requérir des mesures provisionnelles. *Code civil suisse*, *ibidem*.

---

[18] Nel primo *exemplum contrarium*, il verbo *deontico* tedesco *sollen* ha valore *non-normativo*.
[19] Nel primo *exemplum contrarium*, il verbo *deontico* retoromanico *duair* ha valore *non-normativo*, così come ha valore *non-normativo* il suo corrispettivo tedesco *sollen*.
[20] Nel secondo *exemplum contrarium*, il verbo *deontico* tedesco *müssen* ha valore *non-normativo*.

2.3 TERZO *EXEMPLUM CONTRARIUM*. Valore *non-normativo* del verbo *deontico* tedesco *dürfen* in un testo *normativo*.

[14a] Wer bei der Aufmerksamkeit, wie sie nach den Umständen von ihm verlangt werden **DARF**[21] [italiano: ∅; francese: ∅], nicht gutgläubig sein konnte, ist nicht berechtigt, sich auf den guten Glauben zu berufen.
Schweizerisches Zivilgesetzbuch, art. 3, comma 2;

[14b] Italiano:
Nessuno può invocare la propria buona fede quando questa sia incompatibile con l'attenzione che le circostanze permettevano di esigere da lui.
Codice civile svizzero, *ibidem*,

[14c] Francese:
Nul ne peut invoquer sa bonne foi, si elle est incompatible avec l'attention que les circonstances permettaient d'exiger de lui. *Code civil suisse*, *ibidem*.

2.4 QUARTO *EXEMPLUM CONTRARIUM*. Valore *non-normativo* del verbo *deontico* tedesco *können* in un testo *normativo*.

[15a] **KANN**[22] [italiano: ∅; francese: ∅] dem Gesetz keine Vorschrift entnommen werden, so soll das Gericht nach Gewohnheitsrecht und, wo auch ein solches fehlt, nach den Regeln entscheiden, die er als Gesetzgeber aufstellen würde.
Schweizerisches Zivilgesetzbuch, art. 1, comma 2;

[15b] Italiano:
Nei casi non previsti dalla legge il giudice decide secondo la consuetudine e, in difetto di questa, secondo la regola che egli adotterebbe come legislatore.
Codice civile svizzero, *ibidem*,

[15c] Francese:
À défaut d'une disposition légale applicable, le juge prononce selon le droit coutumier et, à défaut d'une coutume, selon les règles qu'il établirait s'il avait à faire acte de législateur. *Code civil suisse*, *ibidem*.

2.5 QUINTO *EXEMPLUM CONTRARIUM*. Valore *non-normativo* del verbo *deontico* tedesco *können* in un testo *normativo*.

[16a] **KANN**[23] [italiano: ***PUÒ***; francese: ∅] nicht bewiesen werden, daß von mehreren gestorbenen Personen die eine oder die andere überlebt habe, so gelten sie als gleichzeitig gestorben. *Schweizerisches Zivilgesetzbuch*, art. 32, comma 2

[16b] Italiano:
Se non **PUÒ**[24] [tedesco: ***KANN***; francese: ∅] essere fornita la prova che di più persone una sia sopravvissuta all'altra, si ritengono morte simultaneamente.
Codice civile svizzero, *ibidem*.

[16c] Francese:
Lorsque plusieurs personnes sont mortes sans qu'il soit possible d'établir si l'une a survécu à l'autre, leur décès est présumé avoir eu lieu au même moment.
Code civil suisse, *ibidem*.

---

[21] Nel terzo *exemplum contrarium*, il verbo *deontico* tedesco *dürfen* ha valore *non-normativo*.
[22] Nel quarto *exemplum contrarium*, il verbo *deontico* tedesco *können* ha valore *non-normativo*.
[23] Nel quinto *exemplum contrarium*, il verbo *deontico* tedesco *können* ha valore *non-normativo*.
[24] Nel quinto *exemplum contrarium*, il verbo *deontico* italiano *potere* ha valore *non-normativo*, così come ha valore *non-normativo* il suo corrispettivo tedesco *können*.

BIBLIOGRAFIA.

COMANDUCCI - GUASTINI
2007 *Analisi e diritto 2006*, a cura di Paolo Comanducci e Riccardo Guastini, Torino, Giappichelli, 2007.

CONTE
2006a al-Malik al-Afḍal, *Modi deontici. 1370*, a cura di Amedeo G[iovanni] Conte, in "Rivista internazionale di Filosofia del diritto" LXXXIII (2006) 479-483.
2006b Amedeo G[iovanni] Conte, *Harem*, in "Fenomenologia e società" XXIX (2006)^4 84-87.
2007a Amedeo G. Conte, *Termini deontici in un antico testo tedesco*, in "Rivista internazionale di Filosofia del diritto" LXXXIV (2007) 109-113.
2007b Amedeo G. Conte, *Fenomeni normativi. Un'indagine non-filosofica*, in COMANDUCCI - GUASTINI 2007, pp. 73-87.

DE VRIES
1962 Jan de Vries, *Altnordisches etymologisches Wörterbuch*, Leiden, E. J. Brill, 1962_2 [1961_1].

TRÀINI
2005 *Uno "Specchio per principi" yemenita: la* Nuzhat aẓ-ẓurafā' wa tuḥfat al-Hulafā' *del sultano rasūlide al-Malik al-Afḍal*, edizione critica con versione italiana annotata a cura di Renato Tràini, in "Atti della Accademia Nazionale dei Lincei" CDII (2005), "Classe di Scienze morali, storiche e filologiche. Memorie" serie IX, volume XIX, fascicolo 2, pp. 225-341. [Il testo arabo è alle pp. 235-284; la versione italiana è alle pp. 285-339.]
2006 *Uno "Specchio per prìncipi" del sultano rasūlide al-Malik al-Afḍal*, in *Storia e cultura dello Yemen in età islamica con particolare riferimento al periodo rasūlide.* Roma, Bardi, 2006 "Accademia Nazionale dei Lincei. Fondazione Leone Caetani", pp. 133-160.

WITTGENSTEIN
[1921]1922/1989 Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*, with an introduction by Bertrand Russell, London - New York, K. Paul, Trench, Trubner & Co - Harcourt, Brace & Company, 1922 [già in "Annalen der Naturphilosophie" (1921)]. Edizione italiana: Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*, introduzione di Bertrand Russell, a cura di Amedeo G. Conte, Torino, G. Einaudi, 1989 "Nuova Universale Einaudi" 196.

WRIGHT
1951 Georg Henrik von Wright, *Deontic Logic*, in "Mind" LX (1951) 1-15.

APPENDICI

MANUEL BARBERA
(Università di Torino)

23. Mapping dei tagset in b.manuel.org / corpora.unito.it. *Tra* guidelines *e prolegomeni.*

0. PREMESSA. I materiali qui presentati da un lato non aspirano a più di essere un aiuto prestato agli utilizzatori di bmanuel.org e corpora.unito.it; da un altro lato, però, sono anche un po' i prolegomeni a quel nostro lavoro sui tagset più volte minacciato (cfr. ad es. qui Barbera ¶ 1, § 3.1), di cui è stato già presentato qui un capitolo (Barbera ¶ 8) recante il quadro teorico generale e la illustrazione del tagset per l'italiano antico, e di cui è in preparazione (da parte di Margarita Borreguero Zuloaga e Marco Tomatis e me medesimo) il capitolo spagnolo (già preannunciato in Barbera 2007 *i.s.*).

Dal secondo punto, soprattutto, discende l'utilità e la latitudine analitica della griglia bibliografica (cfr. § 1), dal primo la limitazione nel mapping presentato nel § 3 ai soli tagset (versioni per TreeTagger) usati, attualmente od in un prossimo futuro, in b.manuel.org e corpora.unito.it, ad esclusione pertanto di altri tagset disegnati esplicitamente per il TreeTagger: penso soprattutto a quello di Achim Stein per lo spagnolo (che è stato la base delle nostre sperimentazioni sullo spagnolo) ed a quello di Marco Baroni per l'italiano (che è la base per il Corpus La Repubblica). Per quanto riguarda il secondo punto, questi dovevano essere (e lo sono stati!) presi ampiamente in considerazione; solo si è scelto di non darne qui conto.

1. BIBLIOGRAFIA RAGIONATA. Per le ragioni esposte nella premessa, una piccola bibliografia ragionata sull'argomento assolve ad un còmpito di utilità al pari della tavola del mapping (cfr. § 3). Ed è forse condensabile in quanto segue (ad esclusione della bibliografia generale sul tagging, comunque ricavabile da Barbera ¶ 8 in questo volume).

Inglese: Santorini 1990 e Marcus - Santorini - Marcinkievicz 1994 (Penn Treebank tagset), Santorini 1991 (rev. di Santorini 1990 per il TreeTagger), TreeTagger Homepage (modifiche al Penn-tagset => Penn-TreeTagger tagset); Teufel 1996 (EAGLES: ELM-EN).

Tedesco: Schiller - Stöckert - Teufel - Thielen 1999 (TreeTagger STTS Tagset); Schiller - Teufel - Thielen 1990, Schiller - Stöckert - Teufel - Thielen 1999 (STTS Tagset); Teufel - Stöckert 1996 (EAGLES: ELM-DE).

Italiano: Stein [2002] (TreeTagger Tagset); Barbera 2007 ¶ 8 (CT-Tagset di antico italiano); Monachini 1996 (EAGLES: ELM-IT); Baroni et alii 2004, p. 1772a, e Baroni [2005] (tagset "La Repubblica").

Francese: Stein 2003 e Stein - Schmid 1995 (TreeTagger Tagset); Rekovski 1996 (EAGLES: ELM-FR).

Spagnolo: Stein [2005] (TreeTagger Tagset); Barbera 2007 *i.s.* (CT-like tagset); Sánchez León 1994 e Sánchez León - Nieto Serrano 1995 (CRATER Tagset); Cabré - Morel - Torner - Vivaldi - Yzaguirre 1998 (IULA Tagset); Brino 2006 (Mapping).

EAGLES: Monachini - Calzolari 1996 (MORPHSYN); Teufel - Stöckert 1996 (ELM-DE); Teufel 1996 (ELM-EN); Rekowski 1996 (ELM-FR); Monachini 1996 (ELM-IT).

2. CENNI METODOLOGICI. Giusto un paio di cenni (l'argomento sarà ripreso in altra sede) ai nostri criteri guida, già applicati nella Ver. 1.2 del tagset spagnolo.

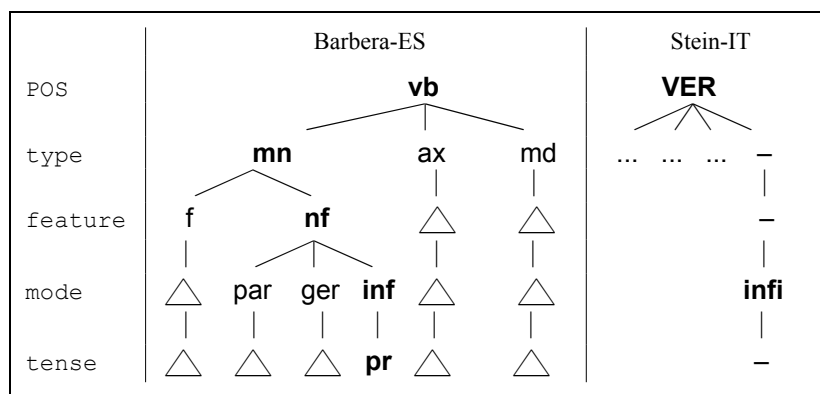
In generale, ai principi di Barbera ¶ 8, che rimangono pienamente validi, se ne sono aggiunti altri due, frutto dell'esperienza cumulata in questi anni.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 373-388.

I principi precedenti, formulati in Barbera ¶ 8, riguardavano questioni sia teoriche (i primi quattro concernevano «i requisiti che un tagset deve soddisfare», *ib*. p. 139) che fattuali (i rimanenti concernevano le «specifiche strutturali generali [cui] deve conformarsi», *ib*. p. 139); li riportiamo qui brevemente: **(1)** consensualità e neutralità (cfr. *ib.* § 1.1), **(2)** adeguatezza descrittiva (cfr. *ib.* § 1.2), **(3)** standardizzazione (cfr. *ib.* § 1.2), **(4)** praticità computazionale (cfr. *ib.* § 1.3), **(5)** tag e *labels* EAGLES-compatibili (corollario di (3), cfr. *ib.* § 2.1), **(6)** ancoramento morfologico (cfr. *ib.* § 2.2), **(7)** struttura tipata (*hierarchy-defining features*: HDF), (cfr. *ib.* § 3 e 3.1), **(8)** evitamento dei *cross-branchings* colla costruzione di gerarchie separate di MSF (*morphosyntactic features*), (cfr. *ib.* § 3.2), **(9)** contenimento dei tag sotto i 70 (corollario di (4), cfr. *ib.* § 4).

I nuovi due principi, **(10)** espansione esplicita di ogni tag gerarchico e **(11)** ottimizzazione ed univocità delle *labels*, sono di livello pratico e sono dei corollari rispettivamente di (7) e di (5), riguardando l'uno la struttura tipata dei tag, e l'altro la scelta dei *labels*.

2.1 ESPANSIONE ESPLICITA DI OGNI TAG GERARCHICO. Il principio (10) cerca di conseguire il massimo sfruttamento della struttura gerarchica con espansione esplicita di ogni tag. In molti (anzi, forse nella più parte dei) tagset la tracciabilità gerarchica è incompleta, ed ovviamente non si possono poi fare query sui nodi rimasti sottointesi.

Se prendiamo come *specimen* la categoria dei verbi, che in EAGLES è una POS con cinque livelli di ramificazione (*branching*), sarà pertanto preferibile avere, ad esempio, un tag composito `vb.ax.nf.inf.pr` che non un più semplice `vb.ax.inf`, rendendo così interrogabile ogni nodo. Si confrontino, infatti, i tag per "infinito di verbo principale" in due tagset rappresentativi:

Tav. 1: La tracciabilità gerarchica di un tag
in tagset a gerarchia esplicita (Barbera, *spagnolo*) ed implicita (Stein, *italiano*).

Nel primo caso si può cercare anche tutti i presenti di qualsiasi modo (`.*.pr.*`), o tutte le forme non finite, incluso partecipi e gerundi (`.*.nf.*`), o tutti i soli presenti di modi non finiti e verbi principali (`.*.mn.nf.+++.pr.*`), laddove nel secondo caso query così mirate non sono possibili.

2.2 OTTIMIZZAZIONE ED UNIVOCITÀ DELLE *LABELS*. Il principio (11) si applica a livello di *labels* anziché di tag, e vuole che le *labels* di ogni tag siano ottimizzate per ottenere query il più possibile univoche. Càpita infatti che per cercare di mantenere il meno "pesanti" possibile[1] le *labels* di un tag, uno incorra in fastidiose omografie (o porzioni significative di essi) tra i nomi dei nodi di gerarchie diverse.

Nella corrente versione del CT-tagset, ad esempio, `ind` è sia un `type` della POS `pd` (pronomi e determinanti), che un `mode` della POS `v` (verbi): quindi una normale query del tipo `.*.ind.*` sarà affatto inefficace, cogliendo sia pronomi che verbi. Seguendo il medesimo ragionamento, è preferibile avere per i verbi una *label* `vb` anziché `v` per evitare che la query `.v.*` colga anche gli avverbi (`adv`) insieme ai verbi.

Di questa esigenza ci siamo avveduti con l'uso della corrente versione del CT-Tagset (ver. 1.3), e ne è in corso la modifica. Analogamente la versione (1.1) del tagset spagnolo presentata in Barbera 2007 *i.s.*, § 3.2, non ne teneva ancora conto, mentre la corrente (1.2) è già stata conformemente rinominata.

La Ver. 1.2 del tagset spagnolo rappresenta così il modello più avanzato nella preparazione dei nostri tagset, cui stiamo lentamente adeguando tutti gli altri.

3. IL MAPPING. Ciò detto, la tavola di mapping tra i tagset è pertanto la seguente (Tav. 2), in cui si sono considerati, accanto ai più elaborati CT-Tagset (italiano antico) e STTS (tedesco), anche il Penn/TT-Tagset (inglese) e due dei tagset (EPADES-like) di Achim Stein (italiano e francese[2]); per lo spagnolo abbiamo invece già dato il nostro, presentato in Barbera 2007 *i.s.*, § 3.2.

Il tagset antico italiano qui impaginato è la attuale Ver. 1.3. che si trova diffusamente illustrata in questo volume in Barbera ¶ 8. Il tagset spagnolo è qui dato nella attuale Ver. 1.2, anziché nella 1.1 presentata in Barbera 2007; le modifiche concernono solo la forma di alcune *labels*, meglio ottimizzate per la query (giusta il principio (11), cfr. *supra* § 2.1).

Salvo quando diversamente indicato (dal corsivo), le glosse fornite sono interlinguistiche e pertanto internazionalmente basate sull'inglese (le eccezioni sono quasi solo per tag (e *labels*) propri ad una sola lingua ed in qualche misura idiosincratici).

Nelle glosse le tonde sono riservate a (parti gerarchiche di) tag non esplicitati in tutti tagset interessati.

Nelle colonne dei tagset, inoltre, le [quadre] rinviano a *labels* di altre POS, le <uncinate> introducono, per chiarezza o quando interlinguisticamente necessario, parti superiori gerarchiche di tag non usate da sole nel tagset in questione.

---

[1] In base alla considerazione che, ovviamente, più una *label* è lunga, più è suscettibile di digitazioni erronee.

[2] Come già detto, il mapping col tagset italiano di Marco Baroni ed antico francese di Achim Stein, pur operazione assai importante ed istruttiva, è stato giocoforza rimandato ad altra sede.

aIT	**ES**	**IT**	**FR**	**EN**	**DE**	
Barbera	Barbera	Stein	Stein	Penn/TT	STTS	
n.c *casa*	nn.cm *casa*	NOM *casa*	NOM *maison*	NN *house*	NN *Haus*	noun, (common)
n.p *Giovanni*	nn.pr *Juan*	NPR *Giovanni*	NAM *Jean*	NP *John*	NE *Johannes*	noun, proper
				NNS *houses*		noun, (common), plural
				NPS *Netherlands*		noun, proper, plural
adj *piccolo*	adj *pequeño*	ADJ *piccolo*	ADJ *petit*	JJ *small*, *happy-go-lucky*	<ADJ>	adjective
					ADJA das *rote* Kleid	adjective, attributive
					ADJD es ist *recht*, er fährt *gekonnt*	adjective, predicative/adverbial
				JJR *smaller*, *better*		adjective, comparative
				JJS *smallest*, *best*		adjective, superlative
adv.g *già*, *non*, *più* *pietosamente*	adv *bien*, *más*, *no* *normalmente*	ADV *già*, *non*, *no*, *solamente*	ADV *ne*, *pas*, *bien* *actuellement*	RB *freely*, *very*, *enough*, *not;* *there*, a melee ensued	ADV *dort*, *heute*, *darum*, *ja*, *nie*, *z.B.*	adverb (general)
				RBR *better*		adverb, comparative
				RBS *best*		adverb, superlative
					PAV es beruft sich *hierauf*	adverbial pro. (*Pronominaladverbien*)
				EX *there* ensued a melee		existential *there*

adv.p *ne*, *ci*, *vi*					<PT>	(adverb), particle
					PTKANT *ja*, *nein*, *danke*, *bitte*	*Antwortpartikel*
					PTKNEG *nicht*	*Negationspartikel*
				RP to run it *up*	PTKVZ er geht *hinein*, steigst du *mit*?	particle; *abgetrennter Verbzusatz*
				TO *to* die or not *to* die	PTKZU ohne *zu* wollen	infinitive particle *to, zu*
					PTKA *am* schönsten, *zu* schnell	*Partikel bei Adjektiv / Adverb*
					KOKOM *als* Taxifahrer	*Vergleichspartikel ohne Satz*
		CON *e*, *ma*, *che*, *se*	KON *et*, *ou*, *que*, *si*			conjunction
con.c *e*, *ma*,	con.c *y*, *o*			CC *and*, *or*, *but*, *plus*	KON *und*, *etweder ... oder*, *denn*	conjunction, coord.
con.s *che*, *se*	con.s *que*, *si*					conjunction, subord.
					KOUS er weiß, *daß* du kommst	conjunction, subord. + sentence (*unterordnende Konj. mit Satz*)
					KOUI sie tun alles, *um* zu überleben	conjunction, subord.+ *zu* + infinitive (*unterordnende Konj. mit Infinitiv*)
				IN *in*, *for*, *at*, [TO]		preposition or conjunction, subord.
adp.pre *a*, *di*	adp.pre *a*, *de*	PRE *a*, *di*	PRP *à*, *de*	[IN]		preposition
					APPR *mit*, *ohne*, *von*, *von* heute an	preposition, *Zirkumposition links*

	adp.pre.art *al*, *del*	PRE:det *al*, *del*	PRP:det *au*, *des*		APPRART *am*, *ans*, *zur*, *zum*	preposition + article
adp.post *-co*					APPO *entlang*, *wegen*	postposition
					APZR von heute *an*, von Rechts *wegen*	*Zirkumposition rechts*
<pd>	<pd>	PRO [sic!] ... *coloro*, *gliel{o}/ne*	PRO [sic!] *quoi*, *qui*		<P>	pronoun
				DT *an*, *the*; *any*, *some*; *that*		determiner
				PDT *all* his marbles, *both* the girls		predeterminer
pd.dem.s *questo*, *ciò*	pd.dem *este*, *ese*	PRO:demo *questo*, *ciò*	PRO:DEM *ce*, *cette*			demonstrative pro./det., (strong)
pd.dem.w *ne*, *ci*, *-vi*						demonstrative pro., weak
				[DT]	PDS *dies* ist ein Buch	demonstrative pro. (*substit. Dem.*)
				[DT]	PDAT *dieses* Buch	demonstrative det. (*attrib. Dem.*)
pd.ind *catuno*, *tutto*, *om*	pd.idf *alguno*, *todo*	PRO:indef *chiunque*, *tutto*	PRO:IND *chacun*, *quelque*			indefinite pro./det.
				[DT]	PIS *nichts*, *keiner* kam	indefinite pro. (*substit. Indef.*)
				[DT]	PIAT *zuviele* Fragen	indefinite det. (*attrib. Indef.*)
					PIDAT *all* die Brüder	indefinite det. (*attrib. Indef.*) + deter.
		PRO:pers *io*, *me*, *mi*	PRO:PER *je*, *on*, *se*, *il*, *la*	PP *I*, *me*, *myself*, *mine*		personal pronoun

pd.per.s.n *io*, *egli*	pd.per.s.n *yo*, *él*					personal pronoun, strong nominative
pd.per.s.o *me*, *sé*, *lui*	pd.per.s.o *mí*, *sí*					personal pronoun, strong oblique
pd.per.w.n *i'*, *-tu*, *e'*						personal pronoun, weak nominative
pd.per.w.o *mi*, *lo*, *gli*	pd.per.w *me*, *se*					personal pronoun, weak oblique
					PPER *ich*, *meiner*	personal pronoun, non-reflexive
		PRO:refl *si*			PRF *mir*, *sich*, *eineinder*	personal pronoun, reflexive
pd.pos.s *mio*, *suo*, *nostro*	pd.pos *mi*, *su*, *nuestro*	PRO:poss *mio*, *suo*, *nostro*				possessive pro./det., (strong)
pd.pos.w *-ma*				POS *-'s*, *-'*		possessive pro./det., weak; possessive ending
			PRO:POS *mien*, *nôtre*	[PP]	PPOSS das ist *meins*	possessive pro. (*substit. Poss.*)
			DET:POS *mon*, *nos*	PRP$ *my*	PPOSAT *seine* Meinung	possessive det. (*attrib. Poss.*)
				WP$ *whose*		possessive wh-pronoun
pd.rel *che*, *cui*, *onde*	pd.rel *que*	PRO:rela *che*, *cui*	PRO:REL *qui*, *que*, *dont*, *ou*			relative pro.-det.
				WP *what*, *who*, *whom*	PRELS derjenige, *welcher*	wh-pronoun
				WDT *which*, *that*	PRELAT der man, *dessen* Hut	wh-determiner
				WRB *how*, *where*, *why* *when* he arrived, I was out	PWAV der Grund, *warum* ich gehe *wo* bist du?	wh-adverb (*adverbiales Interr.*) (~ it. *relativi indefiniti*)

pd.int *che*, *chi*	pd.int *qué*	PRO:inter *quale, come* [sic!]	[PRO], [*:REL] [*.DEM], [*.IND]			interrogative pro.-det.
					PWS *wer* kommt?	interrogative pro. (*substit. Interr.*)
					PWAT *welche* Farbe	interrogative det. (*attrib. Interr.*)
pd.exc *che*, *quanto*	pd.exc *qué*, *quánto*					exclamative (wh) pro./det.
			DET:ART *le*, *un*		ART *der*, *ein*	article
art.d *il*	art.d *el*	DET:def *il*		[DT]		definite article
art.i *un*	art.i *un*	DET:indef *un*		[DT]		indefinite article
		NUM *2, II.* (*due* = ADJ!)	NUM *deux*,*2*,*deuxième*			numeral
num.c *due*	num.cd *dos*, *2*			CD	CARD *drei, 3*	number, cardinal
num.o *secondo*	num.od *tercero*, *duplo*			[JJ]	[ADJ]	number, ordinal
<v>	<vb>	<VER> (incl. *avere*, *dovere* ...)	<VER> (incl. *être*, *avoir*, ...)	VV (incl. *do)* *do* it, you should *do*, to *do*	<VV>	verb, base form
					VVFIN du *gehst*	*finites Verb, voll*
v.m.f.ind.pr *faccio*	vb.mn.fn.ind.pr *sale*	VER:pres *faccio*	VER:pres on *trouve*	[VV]		verb, (main) (fin.) ind. pres.
				VVP I *love*,I *get*		verb, non-3rd pers. sg. pres.
				VVZ he *loves*,he *gets*		verb, 3rd pers. sg. pres.

v.m.f.ind.ipf *facea*	vb.mn.fn.ind.ipf *estaba*	VER:impf *facevo*	VER:impf il *fallait*			verb, (main) (fin.) ind. impf.
v.m.f.ind.pt feci	vb.mn.fn.ind.pt *ocurrió*	VER:remo *feci*	VER:simp je *parlai*	VVD I *loved*, he *got*, I *were*		verb, (main) (fin.) ind. past
v.m.f.ind.ft *farò*	vb.mn.fn.ind.ft *buscaré*	VER:futu *farò*	VER:futu on *trouvera*			verb, (main) (fin.) ind. future
v.m.f.sub.pr *faccia*	vb.mn.fn.sub.pr *cuezan*	VER:cpre *abbiate*, *faccia*	VER:subp pit ce que *soit*	[VV]		verb, (main) (fin.) subj. pres.
v.m.f.sub.ipf *facesse*	vb.mn.fn.sub.ipf *quitaran*	VER:cimp *volessi*, *facesse*	VER:subi q'on *finisse*	[VVD]		verb, (main) (fin.). subj. impf.
	vb.mn.fn.sub.ft *tuviere*					verb, (main) (fin.). subj. future
v.m.f.cnd.pr *faria*	vb.mn.fn.cnd.pr *entraría*	VER:cond *saprebbe*, *farei*	VER:cond nous *devrions*			verb, (main) (fin.).cond. (pres.)
v.m.f.imp.pr *fa'*	vb.mn.fn.imp.pr *vete*	VER:impe *dicci*, *fa'*	VER:impe *sachez*	[VV]	VVIMP *geh* ! , *geht* !	verb, (main) (fin.). impr. (pres.)
v.m.nf.inf.pr *fare*	vb.mn.nf.inf.pr *emplear*	VER:infi *andare*, *andarsene*	VER:infi *faire*	[VV]	VVINF *ankommen*	verb, (main) (n-fin.) inf. (pres.)
					VVIZU *anzukommen*	*Infinitiv mit zu, voll*
		VER:refl:infi [sic!] *pentire*, *vergognare*				verb, (main) reflexive infinitive
v.m.nf.par.pr *vidente*		VER:ppre *viaggiante*	VER:ppre *attendant*	[VVG]		verb, (main) (n-fin.). part. pres.
v.m.nf.par.pt *detto*	vb.mn.nf.par.pt *saludado*	VER:pper *detto*	VER:pper *écrit*	VVN *loved*, *gotten*	VVPP *es wird gesucht*	verb, (main) (n-fin.) part. past
v.m.nf.ger.pr dicendo	vb.mn.nf.ger.pr *hablando*	VER:geru *dicendo*		VVG *loving*, *getting*		verb, (main) (n-fin.) ger. (pres.)
<v.a>	<vb.ax>				<VA> <*haben*, *sein*>	verb, auxiliar

				VB *be* careful, it can *be*, to *be*		verb *to be*, base form
				VH *have* it your way, to *have*		verb *to have*, base form
					VAFIN sie *wären*	*finites Verb, aux*
v.a.f.ind.pr *ha*, *è*	vb.ax.fn.ind.pr *ha*, *es*, *está*			[VB], [VH]		verb, aux. fin. ind. pres.
				VBP I *am*, you *are*		verb *to be*, non-3rd pers. sg. pres
				VHP I *have*, you *have*		verb *to have*, non-3rd pers. sg. pres
				VBZ he *is*		verb *to be*, 3rd pers. sg. pres
				VHZ she *has*		verb *to have*, 3rd pers. sg. pres
v.a.f.ind.ipf *avea*, *era*	vb.ax.fn.ind.ipf *era*					verb, aux. fin. ind. impf.
v.a.f.ind.pt *ebbe*, *fu*	vb.ax.fn.ind.pt *fue*					verb, aux. fin. ind. past
				VBD I *was*, I *were*		verb *to be*, (fin.) ind. past
				VHD I *had*		verb *to have*, (fin.) ind. past
v.a.f.ind.ft *avrà*, *sarà*	vb.ax.fn.ind.ft *será*					verb, aux. fin. ind. future
v.a.f.sub.pr *aggia*, *sia*	vb.ax.fn.sub.pr *sean*			[VB], [VH]		verb, aux. fin. subj. pres.
v.a.f.sub.ipf *avesse*, *fosse*	vb.ax.fn.sub.ipf *hubiese*			[VBD], [VHD]		verb, aux. fin. subj. impf.

	vb.ax.fn.sub.ft *hubiere*					verb, aux. fin. subj. future
v.a.f.cnd.pr *avrebbe*, *seria*	vb.ax.fn.cnd.pr *estaríamos*					verb, aux. fin.cond. (pres.)
v.a.f.imp.pr *abbi*, *si'*	vb.ax.fn.imp.pr *ten*			[VB], [VH]	VAIMP *habt* Geduld !	verb, aux. fin. impr. (pres.)
v.a.nf.inf.pr *avere*, *essere*	vb.ax.nf.inf.pr *ser*			[VB], [VH]	VAINF *sein*	verb, aux. n-fin. inf. (pres.)
v.a.nf.par.pr [*avente*, *essente*]				[VBG], [VHG]		verb, aux. n-fin. part. pres.
v.a.nf.par.pt *avuto*, *stato*	vb.ax.nf.par.pt *sido*				VAPP er ist *geworden*	verb, aux. n-fin. part. past
				VBN *been*		verb *to be*, (n-fin.) part. past
				VHN *had*		verb *to have*, (n-fin.) part. past
v.a.nf.ger.pr *avendo*, *essendo*	vb.ax.nf.ger.pr *siendo*					verb, aux. n-fin. ger. (pres.)
				VBG *being*		verb *to be*, (n-fin.) ger. (pres.)
				VHG *having*		verb *to have*, (n-fin.) ger. (pres.)
<v.d>	<vb.md.>			MD *can*, *could*, *must*, *ought*	<VM> <*können*>	(verb), modal
					VMFIN wir *wollten*	*finites Verb*, *modal*
v.md.f.ind.pr *può*, *vuol*, *dee*	vb.md.fn.ind.pr *puedo*					verb, mod. fin. ind. pres.
v.md.f.ind.ipf *potea*, *volea*	vb.md.fn.ind.ipf *debía*					verb, mod. fin. ind. impf.

v.md.f.ind.pt *poté*, *volle*	vb.md.fn.ind.pt *pudo*					verb, mod. fin. ind. past
v.md.f.ind.ft *potrà*, *vorrà*	vb.md.fn.ind.ft *deberemos*					verb, mod. fin. ind. future
v.md.f.sub.pr *possa*, *voglia*	vb.md.fn.sub.pr *pueda*					verb, mod. fin. subj. pres.
v.md.f.sub.ipf *potesse*, *volesse*	vb.md.fn.sub.ipf *pudiera*					verb, mod. fin. subj. impf.
	vb.md.fn.sub.ft *pudiere*					verb, mod. fin. subj. future
v.md.f.cnd.pr *potrei*, *vorria*	vb.md.fn.cnd.pr *deberían*					verb, mod. fin.cond. (pres.)
v.md.f.imp.pr *dovete*	vb.md.fn.imp.pr *debe*					verb, mod. fin. impr. (pres.)
v.md.nf.inf.pr *potere*, *volere*	vb.md.nf.inf.pr *poder*				VMINF *müssen*	verb, mod. n-fin. inf. (pres.)
v.md.nf.par.pr [*volente*]						verb, mod. n-fin. part. pres.
v.md.nf.par.pt *potuto*, *voluto*	vb.md.nf.par.pt *podido*				VMPP er hat *gewollt*	verb, mod. n-fin. part. past
v.md.nf.ger.pr *potendo*, *volendo*	vb.md.nf.ger.pr *pudiendo*					verb, mod. n-fin. ger. (pres.)
intj *o*, *deh*, *ahi*, *ave*	intj *ay*	INT *beh*, *uh*, *ciao*	INT *hein*, *oui*, *tiens*	UH *oh*, *my*, *please*, *well*, *yes*	ITJ *ach*, *hmh*, *bravo*	interjection
pun.fi . ! ?	pun.fin . ! ?	SENT . ! ?	SENT . ! ?	SENT . ! ?	$. . ! ? ; :	punctuation, final
pun.nfi ; : , ( ) [ ] ' " ...	pun.nfi ,	PON ; : , ( ) [ ] ' " ...	PUN ; : , ( ) [ ] ' ...			punctuation, non-final
				: ; :		punctuation, colon

				, *,*	$, ,	punctuation, comma
					$(([{ "	*sonstige Satzzeichen, satzintern*
				(*({*		punctuation, bracketleft
				) *) }* [not square brackets!]		punctuation, bracketright
				" *" '*		punctuation, quote
			PUN:cit *"*			punctuation, citation
r.for -	r.sy 2+2=4	SYM ***, *+*	SYM *+*, *&*	SYM *π/2*, *[,]*	XY Sie hat *$* 200 verloren	residual, formulae / symbol / *Nichtwort*
				LS *a)*, ·		list item marker
r.frg	r.frg			FW	FM	residual, foreign word
r.abb. *etc~*, *&*, *{DG}*		ABR *pc*, *kg*	ABR *etc.*			residual, abbreviation
					TRUNC *be-* und entladen	*Kompositions-Erstglied* & *Präfix*
r.epe *-e*						residual, Epenthesis

Tav. 2: Mapping tra i vari tagset usati in bmanuel.org e corpora.unito.it.

BIBLIOGRAFIA.

AA. VV.
2004 *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, ELDA, 2004.

ARMSTRONG
1994 *Using Large Corpora*, edited by Susan Armstrong, Cambridge (Mass.) - London (En.), The MIT Pres, 1994 "A Bradford Book", "ACL-MIT Press Series in Computational Linguistics" [= "Computational Linguistics" XIX (1993)[1-2]].

BARBERA
2007 *i.s.* Manuel Barbera, *I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo*, in "Cuadernos de filología italiana" XIV (2007 11-32, in corso di stampa.
¶ 1 Manuel Barbera, *Per la storia di un gruppo di ricerca. Tra bmanuel.org e corpora.unito.it*, in questo volume, pp. 3-20.
¶ 8 Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in questo volume, pp. 135-168.

BRINO
2006 Giovanna Brino, *Problemi morfologici nell'etichettatura morfosintattica dello spagnolo. Strategie e procedure,* Università di Torino, Facoltà di Lingue, Tesi di Laurea, 2004-2005.

BARONI
[2005] [Marco Baroni], [*Italian tagset used for "La Repubblica" corpus in the TreeTagger parameter file*], web page, Stuttgart, IMS website, [2005], `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/ita lian-tagset.txt`.

BARONI et alii
2004 Marco Baroni - Silvia Bernardini - Federica Comastri - Lorenzo Piccioni - Alessandra Volpi - Guy Aston - Marco Mazzoleni, *Introducing the* La Repubblica *Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*, in AA. VV. 2004, pp. 1771-1774, disponibile online alla pagina `http://www.form.unitn.it/~baroni/publications/lrec2004/rep_lrec_2004.pdf`.

CABRÉ - MOREL - TORNER - VIVALDI - YZAGUIRRE
1998 Maria Teresa Cabré - Jordi Morel - Sergi Torner - Jordi Vivaldi - Lluis de Yzaguirre, *El corpus de l'IULA: etiquetaris*, Barcelona, Universitat Pompeu Fabra. Institut Universitari de Lingüistica Aplicada, 1998 "Sèrie Informes" 18; disponibile anche online con la sigla IULA/INF018/98 alla pagina `http://www.iula.upf.es/paps 1ca.htm`.

MARCUS - SANTORINI - MARCINKIEVICZ
1994 Mitchell P. Marcus - Beatrice Santorini - Mary Ann Marcinkievicz, *Building a Large Annotated Corpus of English: The Penn Treebank*, in ARMSTRONG 1994, pp. 273-290. Disponibile online dalla homepage del PennTreebank al link `ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz`.

MONACHINI
1996 Monica Monachini, *ELM-IT: EAGLES Specifications for Italian Morphosyntax - Lexicon Specifications and Classification Guidelines*, Pisa, EAGLES Document EAG-CLWG-ELM-IT/F, May 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`.

MONACHINI - CALZOLARI
1996 Monica Monachini - Nicoletta Calzolari, S*ynopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Application to European Languages*, Pisa, EAGLES Document EAG-CLWG-MORPHSYN/R, May 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`.

REKOWSKI
1996 Ursula von Rekowski, *Specifications for French Morphosyntax - (ELM-FR)*, Paris, EAGLES Document EAG-CLWG-ELM-FR/F, 31st Aug. 1996. Disponibile online alla pagina: `http://www.ilc.cnr.it/EAGLES/browse.html`.

SÁNCHEZ LEÓN
1994 Fernando Sánchez León, *Spanish tagset for the CRATER Project*, PDF file, Doc. id. CRATER/WP6/FR1, March 7, 1994; disponibile online come Arxiv eprint *arXiv:cmp-lg/9406023* ver. 1 alla pagina `http://arxiv.org/abs/cmp-lg/9406023`.

SÁNCHEZ LEÓN - NIETO SERRANO
1995 Fernando Sánchez León - Amalio F. Nieto Serrano, *Development of a Spanish Version of the Xerox Tagger*, PDF file, Doc. id. CRATER/WP6/FR1, May 19, 1995; disponibile online come Arxiv eprint alla pagina `http://arxiv.org/abs/cmp-lg/9505035`.

SANTORINI
1990/1 Beatrice Santorini, *Part-of-speech Tagging Guidelines for the Penn Treebank Project*, Technical report MS-CIS-90-47, University of Pennsylvania - Department of Computer and Information Science, 1990. *3rd Revision, 2nd Printing, June 1990* è disponibile online dalla homepage del PennTreebank `ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`; la *Rev. 1991 March 15* è disponibile dalla homepage del Treetagger al link `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps`.

SCHILLER - STÖCKERT - TEUFEL - THIELEN
1999 Anne Schiller - Simone Teufel - Christine Stöckert - Christine Thielen, *Guidelines für das Tagging Deutscher Textkorpora mit STTS. (Kleines und großes Tagset)*, Technical report, IMS and SfS, disponibile online alla pagina `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSe ts/stts-1999.ps.gz`.

SCHILLER - TEUFEL - THIELEN
1995 Anne Schiller - Simone Teufel - Christine Thielen, *Guidelines für das Tagging Deutscher Textkorpora mit STTS*, IMS and SfS, Draft 26 September 1995, disponibile online a `http://www.sfs.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz`.

STEIN
[2002] Achim Stein, *Italian tagset used in the TreeTagger parameter file*, web page, Stuttgart, IMS website, [Ver. 22 April 2002], `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/italian-tagset.txt`.

2003 Achim Stein, *French TreeTagger Part-of-Speech Tags*, html page, Stuttgart, IMS website, April 2003, http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html.

[2005] [Achim Stein], [*Spanish tagset used in the TreeTagger parameter file*], web page, Stuttgart, IMS website, [11 August 2005], http://www.ims.uni-stuttgart.de/ftp/pub/corpora/italian-tagset.txt.

STEIN - SCHMID

1995 Achim Stein - Helmut Schmid, *Étiquetage morphologique de textes français avec un arbre de décisions*, e-paper, Stuttgart, IMS website, 1995, http://www.uni-stuttgart.de/lingrom/stein/pubs/steins95.ps.gz o http://www.uni-stuttgart.de/lingrom/stein/pubs/pdf/steins95.pdf.

TEUFEL

1996 Simone Teufel, *ELM-EN. EAGLES Specifications for English Morphosyntax. Draft Version*, Stuttgart, EAGLES Document, July, 31 1996. Disponibile online alla pagina: http://www.ilc.cnr.it/EAGLES/browse.html.

TEUFEL - STÖCKERT

1996 Simone Teufel - Christine Stöckert, *ELM-DE. EAGLES Specification for German Morphosyntax. Lexicon Specification and Classification Guidelines*, Stuttgart, EAGLES Document EAG-CLWG-ELM-DE/F, März 1996. Disponibile online alla pagina: http://www.ilc.cnr.it/EAGLES/browse.html.

CORPORA E SITI DI RIFERIMENTO.

bmanuel.org	http://www.bmanuel.org
corpora.unito.it	http://www.corpora.unito.it/.
Corpus Taurinense	http://www.bmanuel.org/projects/ct-HOME.html
CWB	http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
ELWIS	http://www.sfs.uni-tuebingen.de/Elwis/
IMS Stuttgart	http://www.ims.uni-stuttgart.de
“La Repubblica” C.	http://sslmit.unibo.it/repubblica
NUNC	http://www.bmanuel.org/projects/ng-HOME.html
PennTreebank	http://www.cis.upenn.edu/~treebank/
Tree Tagger	http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

MANUEL BARBERA - ELISA CORINO - CRISTINA ONESTI
(Università di Torino)

24. Indice analitico[1].

[1] I rinvii sono alla pagina del volume; in grassetto i riferimenti salienti; in corsivo quelli contenenti un riferimento web o bibliografico di base. Di norma le occorrenze di ogni type sono segnalate fuse una pagina alla volta, tranne che quando pertinenti a categorie diverse e per gli esempi linguistici. I rinvii interni (così come i titoli ed i titoli correnti) non sono indicizzati. Le espressioni indicizzate, ovviamente, lo sono solo nel loro valore proprio o specialistico e non in valori secondari, generici o metaforici (quindi, ad esempio, solo le occorrenze di *occorrenza* 'token, *instance*' e non 'bisogno, esigenza'). Nell'indice la virgola "," funge da operatore di inversione, la barra verticale "|" da operatore *and-or*, e la barra obliqua "/" da funtore gerarchico.

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 389-415.

MAURO COSTANTINO
(Università di Torino)

25. Indice dei nomi[1].

[1] I numeri di pagina citati in corsivo rimandano ai riferimenti bibliografici; per le citazioni esplicite dagli autori si è scelto di adottare il sottolineato; un asterisco precede il riferimento ad una nota; il riferimento più saliente, dove ritenuto opportuno, è evidenziato in grassetto. [M. C.]

Come d’uso: la virgola è l’operatore di inversione; i “de” ecc. non sono alfabetizzati in spagnolo e francese, ma lo sono in italiano, e così i “van” non sono alfabetizzati in tedesco ma lo sono in olandese; cognome e nome non sono separati dalla virgola in lingue (*e.g.* ungherese e giapponese) dove costituiscono l’ordine normale (e pertanto non richiedono un operatore di inversione); i nomi storici (*e.g.* Dante Alighieri, Arnaut Daniel) sono alfabetizzati sotto il nome anziché il cognome. In casi dubbi ha fatto fede il catalogo della New York Public Library. [M. B.]

Corpora e linguistica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 417-427.

MANUEL BARBERA
(Università di Torino)

27. Indice dettagliato.

Corpora e lingustica in rete, a cura di Manuel Barbera, Elisa Corino e Cristina Onesti, Perugia, Guerra Edizioni, 2007, pp. 429-438.

Manuel Barbera
è ricercatore presso il Dipartimento di scienze letterarie e filologiche dell'Università di Torino. Filologo e linguista storico e computazionale, si è occupato di corpus linguistics, uralistica, altaistica, romanistica, italianistica, metricologia, critica testuale, paleografia, font editing, lessicografia e didattica delle lingue.

Elisa Corino
svolge attività di ricerca presso il Dipartimento di scienze letterarie e filologiche dell'Università di Torino, dove si occupa di germanistica e di didattica dell'italiano come L2, coordinando tra l'altro il corpus di apprendenti VALICO ed il suo monitor corpus VINCA di italiani madrelingua.

Cristina Onesti
svolge attività di ricerca presso il Dipartimento di scienze letterarie e filologiche dell'Università di Torino. I suoi interessi sono rivolti principalmente alla linguistica tedesca, alla magiaristica ed alla glottodidattica. Coadiuva alla realizzazione del corpus giuridico Jus Jurium.

corpora e linguistica in rete

a cura di: Manuel Barbera, Elisa Corino e Cristina Onesti

Questo volume, attraverso i contributi del gruppo di ricerca torinese e dei suoi collaboratori in Danimarca, Germania, Spagna e Svizzera, realizza con analisi metalinguistiche ed applicate un percorso organico ed originale nella linguistica dei corpora.
Gli aspetti legali dei corpora ricevono per la prima volta una trattazione esauriente, con proposte innovative e risolutive, concertandosi ad una approfondita messa a punto teorica dei concetti cardine della disciplina (corpus, token, markup, POS-tag), senza trascurare aspetti più tecnici come la rappresentazione computazionale dei corpora e la costruzione e standardizzazione dei tagset.
Le potenzialità offerte dai nuovi tipi testuali presenti in rete, specie dai newsgroup, sono esaminate con particolare attenzione. I case studies spaziano dall'àmbito lessicografico, alle implicazioni grammaticografiche ed all'applicazione glottodidattica dei materiali a disposizione, con speciale riguardo al dominio testuale, di solito poco rappresentato nella linguistica dei corpora, articolandosi su più lingue (antico italiano, danese, francese, inglese, spagnolo, tedesco) con al centro l'italiano contemporaneo.

salt&pepper_perugia

ISBN 978-88-557-0041-2
9 788855 700412